RADIO ENGINEERING HANDBOOK

KEITH HENNEY, EDITOR-IN-CHIEF

Fellow, The Institute of Radio Engineers,
Co-author, "Principles of Radio," "Electron Tubes in Industry", Consulting Editor, "Nucleonics"

FIFTH EDITION

BOOK COMPANY, INC.

NEW YORK TORONTO LONDON

I

Copyright ©

Copyright, 1933, 1935, 1
in the United States
thereof, may not be re

Library ... Catalog Card Number: 58-11171

PREFACE

In preparing new material and revising existing material for this Fifth tion, the same guiding principles have been followed as in the First ition in 1933. Throughout, the endeavor has been to create a compre-nsive working manual and to compile in a single volume concise formation on each of the branches of communication engineering.

science with as many variable parameters as electronics, the only daries are those of the imagination of the scientists and engineers. ach decade brings forth totally new concepts and new tools. The riod from 1950, when the Fourth Edition was published, to 1959 has n no exception. There has been no stagnation; no settling down; no ange from dynamic progress.

This edition, therefore, contains much that appeared only in brief form, at all, in the previous editions. All chapters have been brought up to ate; seven of them have been totally rewritten In addition seven apters are totally new And because there is no longer much difference etween the tools employed by telegraph and telephone engineers and ose used by radio men, this volume now has chapters devoted to wire-ided communication. Thus this Fifth Edition becomes a "com-unications" handbook and not one concerned solely with space-guided formation transmission. The original name, however, has been tained

While there is much of what may be called fundamental background in is book, the emphasis has been on working practice rather than theory general concept of the purpose of a "handbook."

The engineer will find here many man-hours of effort compiled in the rm of tables and curves or converted into concise English by the engi-rs, physicists, and teachers who have aided the editor in preparing this w edition To these authors the editor wishes to express his very great

apter 9. *Sem*

KEITH HENNEY

CONTENTS

CHAPTER 1

BASIS OF RADIO-COMMUNICATION ENGINEERING

By Beverly Dudley[1]

1. Introduction. Radio-communication engineering may be defined as the art and science by which materials and natural phenomena are utilized in the design, construction, and operation of apparatus to convey information or to effect remote control between two or more points through the use of electromagnetic waves.

Man has no knowledge of, and no contact with, his environment, except as through his senses. Therefore, if man is to benefit from radio-communication engineering, information must be conveyed to him by variation of forms of energy that will stimulate or affect one or more of his senses. Moreover, if information is to be conveyed, some prearranged and generally understood sequence of symbols or signals—some form of language, however rudimentary, and generally expressed as a single-valued function of time—must be developed and understood by the parties involved.

Design and construction of instrumentalities for effective radio communication demand the proper use of material substances and that the forces of electricity and magnetism be manipulated to produce electromagnetic waves, modified to contain or represent the information to be conveyed, simultaneously possessing characteristics to enable their passage through the transmission medium. Since there appears to be no economical or practical possibility that man can control the characteristics of the earth's atmosphere, successful radio communication (especially that conducted over long distances) requires that the electromagnetic-propagation characteristics of the earth's atmosphere be observed and studied so that natural phenomena may be utilized most advantageously.

Modern radio communication therefore rests upon the five fundamental bases:

1. The physical, physiological, and psychological foundations of man's sense perceptions, especially those of seeing and hearing
2. The analysis, formulation, and manipulation of intelligence expressed as a single-valued function of time in a form suitable for transmission between two or more points
3. The fundamental structure of matter, and particularly the manner in which the properties of material substances depend upon, or are influenced by, the composition and arrangement of elementary particles of which they are composed
4. The fundamentals of those branches of physics (primarily electromagnetism) by which electromagnetic waves are generated, modulated, launched at a transmitter, and intercepted at a receiving point where the intelligence imparted in the modulation process is extracted by a process of demodulation or detection
5. The propagation characteristics of the medium through which the electromagnetic waves travel from the transmitting end to the receiving point

Because the scientist and engineer are concerned with quantitative relationships between the numerous entities and variables encountered in the five fundamental bases enumerated above, mathematics, "the science of treating of the exact relations existing between quantities or magnitudes and operations, and of methods by which, in accordance with these relations, quantities sought are deducible from others known or supposed," may be regarded as a sixth fundamental basis of radio engineering. In addition to all the elementary branches of mathematics, radio engineers may be called upon to make use of such intermediate and advanced mathematical subjects as the

[1] Massachusetts Institute of Technology.

differential and integral calculus, ordinary and partial differential equations, vector analysis, functions of a real and complex variable, statistics and probability theory, symbolic and operational methods such as Fourier and Laplace transformations, the calculus of variation, and generalized harmonic analysis.

No single book can hope to provide adequate presentation of all the data that a radio engineer may need in the pursuit of his profession, and this first chapter, concerned primarily with developing an outline of the five fundamental bases enumerated above, will err from the standpoint of completeness because of the necessity for brevity and conciseness. Numerous topics must be omitted completely, and others have not yet reached the point where they lend themselves to definitive and concise handbook treatment. Fuller treatment may be found in the many references listed.

PRINCIPLES OF ELECTRICAL COMMUNICATION

2. Signals. Communication is carried on by the successive selection and arrangement of prearranged or understood symbols. The proper selection and sequential arrangement of such symbols or signals at one part of the communication system provide the means by which intelligence (that cannot be completely predicted at some other point of the system) is conveyed, in accordance with the meanings attached to, or the information represented by, the symbols. Since a sequence of symbols is required to convey information, one of the important characteristic properties of intelligence is that it can be expressed as a single-valued function of time.

Each separate and different symbol must vary in some respect from all others, and it must also convey a separate and distinct meaning. Each different symbol must be unique; it must represent or express a unique meaning, and it must be expressible as a unique function of time. Hence a group of different signals or symbols represents a variety of unique time functions; it also represents a variety of unique frequency functions, since time and frequency are inversely related. A group of signals or symbols representing intelligence may therefore be represented by a band of signal frequencies.

These statements are valid for all kinds of communication systems, no matter what means may be employed to convey intelligence between two or more points. These are important results for the engineer, for they make possible the effective and economical transmission of information by electrical means.

Another important characteristic of intelligence or information is its unpredictability. *No new information is conveyed by any operation that is completely predictable.* This point, which has received considerable attention since the Second World War, is one of the cornerstones of modern communication or information theory.

3. Speed of Electrical Communication. The speed with which signals may be transmitted from one point and reproduced at another distant point is of great practical importance. Because electromagnetic phenomena may be propagated between two points with speeds that can, and usually do, approach the velocity of light (186,300 miles per second), electrical methods of communication have replaced, almost entirely, other means of communication over appreciable distances.

4. The Carrier and the Transmission Medium. For satisfactory communication to be effected between two or more points by means of electrical energy, a transmission path must exist between those points. When a suitable path is provided, electrical energy then serves as the vehicle, or *carrier*, by which communication may be effected. In wire-communication systems, electric current or voltage serves as the carrier. In ordinary local telephone or telegraph systems, the carrier current may be unidirectional and unvarying; in other words, it is a direct current of zero frequency. In more complex systems, the carrier current may oscillate periodically and is then characterized by the frequency of oscillation. In radio systems, the carrier wave is comprised of periodic oscillations of the electromagnetic field which determine the frequency of the carrier wave. Electromagnetic waves used as the carrier in radio systems can be produced conveniently and economically only at frequencies of 10,000 cps or more, and usually frequencies very much greater than these are employed.

Of itself, steady-state electrical energy in the appropriate form for passage through

the transmission medium is insufficient to permit significant communication to take place. Electrical energy does not produce sensory perceptions directly. But even if this possibility did exist, it would still be necessary to impart to the carrier the intelligence contained in, or represented by, the signal, for the unmodulated carrier alone contains only trivial information.

5. Electrosensory Transducers. Because electrical energy does not directly produce sensory perceptions or conversely, all systems of electrical communication require: (1) some means for converting sense perceptions into electrical energy at one point and (2) some means at a receiving point for reconverting the received electrical energy into a form capable of stimulating or exciting one or more of the senses. Thus electrosensory transducers—or equivalent substitutes—are required at those points at which intelligence is to be imparted to or extracted from the communication system. These transducers are almost always—but not necessarily—complementarily related, so that the reproduced form of intelligence is of the same kind as the original form of intelligence.

By an electrosensory transducer at the transmitter, sense-perceptory energy corresponding to the selection and sequential arrangement of symbols and representing the intelligence to be conveyed is converted into variations of electrical energy as a single-valued function of time. Such energy conversion occurs at the signal frequency, or band of frequencies, and represents the signal information. Hence the signal variations of energy of the system must be related to, or must correspond to, the frequency band represented by the selection and sequential arrangement of symbols; this must also be true at each and every point of the transmission system.

6. Functions of Signal and Carrier. If the carrier is to convey information through the transmission medium, some property of the carrier must be modulated or varied in accordance with the signal. In a modulated carrier, the carrier wave enables the electrical energy to be passed through the transmission medium, whereas the modulation at signal frequencies imparts to the carrier the information to be communicated.

When the modulated carrier wave has passed through the transmission medium and is intercepted at a receiving point, it is passed through a suitable demodulating device, or detector, so that the signal frequencies may be separated from the carrier. The signal frequencies then actuate an electrosensory transducer so that the signal information is reconverted into an appropriate sequence and arrangement of sense-perceptory symbols. The reproduced symbols should be sufficiently like those initially impressed upon the system that they will be recognized for their intended meaning.

7. Channel Selection. For most kinds of radio systems, the carrier wave is generated as a harmonically varying time function of voltage, current, or field intensity. Thus, the instantaneous value of the carrier may be $i = I_c \cos(\omega t + \phi)$, where I_c is the maximum value of the harmonic variation, $\omega = 2\pi f$ is the angular frequency in radians per second if f is the cyclic frequency in cycles per second (cps), and ϕ is the phase angle which determines the value of the function for time $t = 0$. For such a wave, the carrier is characterized by a single, constant, and fixed frequency, $f = \omega/2\pi$ cps. When a constant-frequency carrier is modulated by a band of signal frequencies, the carrier frequency is no longer constant and fixed but is broadened into a band of frequencies, the spectrum of the modulated wave usually extending above and below the carrier frequency by an amount depending upon the band of signal frequencies with which the carrier is modulated. The frequency spectrum occupied by a carrier and its signal side bands is commonly called a *communication channel.*

In the process of modulation, information is shifted from the band of signal frequencies to a higher (occasionally to a lower) part of the frequency spectrum adjacent to the carrier frequency. The modulated wave has information-bearing side bands with essentially the same propagation characteristics as the carrier wave. Through the use of equipment of such spectral selectivity as will accept the carrier frequency and the side-band frequencies, the information contained in the band of signal frequencies becomes available at the receiving point, although shifted in the spectrum.

More than one modulated carrier, each of which contains a different message or

signal, can be transmitted simultaneously through the same transmission medium. Provided each carrier wave with its associated side bands occupies a frequency channel or a portion of the frequency spectrum devoid of frequency components of some other modulated carrier wave (with its associated side bands), it is possible to separate the modulated carriers on a frequency-selective basis, and to obtain the signal information imparted to each carrier, free from interference from the other modulated carriers in the medium.

The success of such channel-selection techniques usually requires that each modulated carrier operate within its assigned channel or frequency spectrum. For the most part, the frequencies of each separate modulated carrier must lie within the assigned frequency channel, the frequencies of adjacent channels should not overlap, and the frequency selectivity of the receiving apparatus must be such as to accept substantially all frequency components of one channel and exclude the frequency components of neighboring channels.

8. General Problem of Electrical Communication. The general problem is to make possible, as well as economically feasible, all the essential operations of the communication system. Moreover, the system as a whole must be so organized and correlated that the sense perceptions reproduced at the receiving points are essentially the same as those initiated at, and impressed upon, the system at the transmitting point. The means by which the various individual steps are accomplished depend upon: (1) the type or mode of communication, (2) the kind or type of information to be communicated, and (3) the area or location of the receiving points relative to that of the transmitting point or points.

9. Fundamental Steps in Electrical Communication Systems. No matter how simple or how complex the communication system may be, the essential steps may be divided into four distinct operations:

1. The *formulation* of intelligence—expressed by means of signals or symbols—in a manner capable of producing subjective sensations or of effecting remote-control operations. This step is common to all communication systems, whether electrical or not. In radio systems, the most representative types of signals are the dot-dash signals of the Morse code in telegraphy, voice or music signals in telephony, and the scanning signals for picture transmission.
2. The *conversion* of the intelligence into electrical energy of such properties as to be capable of conveying information through the selected transmission medium. In systems of electrical communication, the band of signal frequencies required is superimposed upon the carrier frequency in a process of modulation. Each set of side bands may be identified and separated from all others on a carrier-frequency-selection basis, usually by assigning a separate carrier-frequency band or channel for each different kind of service, station, or message.
3. The *transmission* of the electrical representation of the intelligence from the sending end and its reception at the receiving end of the system.
4. The *retranslation* of the received electrical representation of the information into a form of energy that is capable of affecting the senses of the recipient to produce a subjective sensation essentially the same as that of which the original intelligence formulation was capable, or of effecting some intended control operation.

10. Functional Elements. These considerations permit the establishment of fundamental sequential steps and concepts. Let i represent the intelligence, e the electrical form of the energy corresponding to this intelligence, t designate time, and let the subscript 1 designate those functions or operations occurring at the transmitting portion of the system and the subscript 2 operations occurring at the receiving end. With this notation, the functional operations in any electrical communication system may be expressed, symbolically, as

$$i_1(t) \rightarrow e_1(t) \rightarrow e_2(t) \rightarrow i_2(t)$$

in which the arrows indicate the direction in which the operations proceed. This form of representation expresses the four fundamental energy conversions enumerated above, for an ideal system free from noise, interference, or distortion.

11. Correlation of Transmitting and Receiving Functions. Aside from the speed of transmission, two characteristic features of the over-all system are primarily responsible for its effectiveness and versatility.

The first is the manner in which the signal or information to be conveyed is impressed upon the carrier and extracted from it. In some systems the modulating and demodulating methods are sufficiently simple and independent of each other that (except that both must be responsive to the same carrier-frequency channel) the transmitter and receiver may be designed separately and still be capable of operating together as a unified system. In more sophisticated techniques the modulation and demodulation functions must be correlated and interrelated to function as a unified whole. This is especially true when operations at the receiving end must be synchronized with those at the transmitting end. For such systems the transmitter and receiver cannot be designed as independent and unrelated units if success is to be achieved; from the beginning, both must be planned as integrated elements.

The second characteristic feature is the manner in which the wave emitter at the transmitting end and the wave interceptor at the receiving end may be operated conjointly to determine the transmission path between the two and to take advantage of its propagation properties. In some of the simpler radio systems in which transmitting and receiving apparatus may be designed separately, the wave emitter and the wave interceptor may be randomly related. In more complex and efficient systems there is a tendency to employ directional properties of transmitting and receiving antennas to propagate electromagnetic energy in a reasonably well defined transmission path. The emitter and interceptor may be oriented in azimuth and in elevation. In some kinds of service the directional properties of the radiating and receiving antennas may also be altered at different times, and in this case time synchronization of transmitting- and receiving-antenna structures is required. When this is the case, the transmitter and receiver can no longer be designed or operated without regard to performance at the other end of the system.

The interrelationships between transmitting and receiving operations with respect to modulation methods, synchronization of operations, and propagation path may be illustrated by considering the over-all performance and requirements of a number of different kinds of typical radio systems.

12. Radio Telegraphy. In these systems, the different letters that spell out words are resolved into elementary step functions, represented by dot-and-dash impulses, suitably spaced in time. Initially, therefore, telegraph signals are essentially created as a sequence of single-valued time functions. Each symbol for a letter, numeral, or mark differs from all others by a distinctive and unique combination of the active and inactive periods of each dot-dash-space combination of the code by which the letters are expressed. By keying, some property of the carrier wave is modified or modulated in accordance with the intelligence. The keying operation imparts information to the carrier and therefore it is, functionally, equivalent to an electrosensory transducer. After the modulated or keyed carrier wave passes through the system, the signal information is extracted in such a way as to actuate a loud-speaker, sounder, buzzer, recorder, meter, teletypewriter, or other electrosensory transducer to make the dot-dash symbols visible or audible. The receiving equipment must be tuned to the spectrum occupied by the carrier and its side bands, and it must have suitable demodulating apparatus, but otherwise the design of the receiver can be quite independent of the design of the transmitter, at least so far as concerns radio-telegraph systems employing manual keying.

13. Radiotelephony. In telephony, intelligence is conveyed by uttering, syllable by syllable, the words of sentences making up the intelligence. For speech, music, or other sounds, only one value of sound pressure is produced at any instant of time so that the original form of the signal is created as a single-valued time function. By a microphone and suitable modulating apparatus, some property of the carrier is modulated as a single-valued function of time in accordance with the instantaneous sound pressure of the original signal. After passage through the medium, the intelligence is

extracted from the received modulated wave by a demodulator, and the demodulated signal actuates a headset or loud-speaker to reconstruct the original signal. Again, the receiver must be tuned to the carrier channel and it must have suitable demodulating apparatus, but otherwise receiver and transmitter can be designed and operated independently.

14. Picture Transmission. Visual-image transmission, on the other hand, presents problems not encountered in sound transmission. A picture or visual image is seen, subjectively, by each eye as a two-dimensional area. To be suitable for transmission by electrical means, the images must be resolved into a single-valued function of time. Usually the problem is simplified somewhat by neglecting the fact that the two eyes see slightly different views as a result of their separation by the interpupillary distance, and by selecting for transmission a single view as might be seen by either eye alone.

As a first step, the scene is divided into an orderly array of comparatively minute areas or elements, each of which may be considered to be of uniform visual characteristics. The picture elements are then converted into single-valued time functions through a process of scanning and energy conversion carried out simultaneously. In the scanning process, the picture elements are progressively and systematically selected according to a prearranged sequence which automatically effects the necessary space-to-time conversion. As each picture element is successively selected, it is examined by an electrovisual transducer whose electrical output is related to some visual property of the picture element.

To accomplish the transmission of monochrome pictures by electrical means, some property of the signal wave may be varied in accordance with the visual luminosity of each picture element as it is scanned. The electrical transmission of colored images, on the other hand, requires that the complete picture or image be analyzed, on a color basis, into partial images. Each partial image contains or represents a different chromatic property of the complete scene, in such a way that the combination of all partial images, each properly chromatized, reconstructs the original color image. Each partial image must then be analyzed into picture elements which are systematically scanned, converted into electrical energy, and transmitted as single-valued functions of time.

For the transmission of still pictures, the time required for the scanning operation is not usually of critical importance, and it may extend over a number of minutes or even longer time intervals. For the transmission of visual images in which motion is to be portrayed, a number of additional complications are involved, including—primarily for psychological reasons—a maximum time during which the transmission of any given scene or image must be completed.

The visual effect of apparent motion is achieved through use of a series of instantaneous "still" views, known as *frames* or *fields*. Each frame or field thus represents the image at a given instant, and a series of such frames or fields is made with successive frames exposed in sequence for comparatively short intervals of time. When a series of such frames is properly recorded and subsequently viewed in the same order in which they were exposed, and when the successive frames follow one another with sufficient rapidity to minimize objectionable flicker, the subjective effect is that of apparent motion. The apparent speed of motion of moving objects when successive frames are viewed will be the same as the actual speed of motion of the object photographed when the speed at which successive frames is viewed is the same as the speed with which successive frames were exposed.

To produce the effect of apparent motion without objectionable flicker, the viewing rate is usually about 24 to 30 frames per sec or more. In the transmission of monochrome pictures, each frame or field is transmitted in about $\frac{1}{30}$ sec or less. If the pictures are in color, each frame or field must be resolved into its partial color images. There are usually three partial images per frame and since all partial images of the same frame or field are to be transmitted in about $\frac{1}{30}$ sec, each partial image should be transmitted in about $\frac{1}{100}$ sec. Although variations of this method are possible,

the outline given here is sufficient to indicate the general requirements for transmitting still or motion pictures, either in monochrome or in color, by electrical means.

At the receiving end of a picture-transmission system, the single-valued time function representing the scanned image must be reassembled or reconstructed into a spatial display or array of monochrome or color picture elements in such a way as to produce an image closely resembling the original. This operation requires that certain functions at the receiving end of the system be synchronized with the corresponding complementary functions performed at the transmitting end.

If the system is to operate satisfactorily as a unit: (1) the order in which elements are reconstructed for each frame at the receiving end must be identical to the order in which the image is analyzed at the transmitter; (2) the rate of re-creating elements into a picture at the receiver must equal the rate at which the image is analyzed into picture elements at the transmitter; (3) the order in which partial color images are reconstructed at the receiving end must be the same as that in which the image was analyzed into partial color images (primary colors, or chromaticity, hue, and saturation) at the transmitting end; and (4) the relative positions of the re-created picture elements in each frame of the received picture must be the same as the relative positions of the elements into which the image is analyzed at the transmitting end. Thus, all operations in the synthesis of an image at the receiver must be synchronized with complementary operations in the analysis of the image at the transmitter.

Transmission and reception of pictures require that the transmitting and receiving apparatus must be designed and operated as a consistent, interdependent, synchronized system as a whole.

15. Broadcasting. To reach as large a potential audience as possible, broadcast transmitters are usually located in or near large concentrations of population. The radio signals are allowed to radiate from the transmitting antenna more or less uniformly in all directions along the surface of the earth, although some directivity is often provided to minimize interference from other broadcasting stations, or to maximize the potential audience. Receiving systems for broadcast reception are not especially directive. So long as signals from the broadcast transmitter can be received more or less reliably, there is not much need to correlate or integrate the wave-emitting and wave-intercepting apparatus in broadcast systems.

16. Point-to-point Communication. For point-to-point communication, the objective is usually that of transmitting the largest possible signal most reliably from a given point to some other given point and so there is usually distinct advantage in not having the signal available for interception at intermediate points. A high degree of directivity in emitting or propagating the signal is usually desirable; a high degree of sensitivity of the receiving apparatus in the direction from which the signals arrive is also usually desirable.

The directivity or nondirectivity with which radio signals are propagated and intercepted depends primarily upon the design of the antennas at the transmitter and at the receiver. It also depends upon the propagation characteristics of the earth's surface atmosphere. The degree of directivity that can be achieved practically and economically is roughly proportional to the carrier frequency. Directional-transmission properties are closely associated with the propagation characteristics of the radio spectrum, as indicated by ultrahigh-frequency systems employing forward scatter, for example.

For the directive properties of a radio system to be utilized most effectively, it is necessary that directive antennas at both ends of the circuit be properly correlated in azimuth, and it may also be required that they be correlated in elevation angle, especially if directive communication is to be carried on over long distances.

17. Radio Direction-finding Systems. Direction-finding systems are designed so that the location of a radio transmitter can be determined from the radio signals of known modulation and radiation characteristics. One method (which does not require special equipment at the transmitter) requires that fixed receiving stations with variable directive antenna systems simultaneously take bearings on signals from

the transmitter whose location is to be ascertained. By methods of triangulation, directional bearings from two or more d-f stations are plotted on a map. The point where such triangulation "fixes" intersect determines the location of the station whose position is desired. In other systems, such as loran, for example, signals of known directivity or of known modulation or timing characteristics are transmitted from fixed stations whose positions are precisely known. By means of suitable receiving equipment, operators can determine their position without depending upon fixed stations to determine location "fixes" by triangulation or similar methods.

Radio beams of known azimuthal directivity and elevation are employed in a number of instrumental landing systems. Signals of suitable directivity pattern are radiated by fixed stations so that, by means of appropriate receiving apparatus and indicators, airplanes may land without visual reference to the ground. Directive beams for course determination may also be modulated to indicate "on-course" or "off-course" information or to convey other information besides that of directivity and distance from the transmitting station.

18. Radar Systems. Radio detection and ranging techniques make it possible to detect the presence and determine the character of material objects, and to determine their direction, distance, and elevation. Although radar signals may be generated as continuous waves, or may be frequency- or pulse-modulated, pulse modulation is now almost universally employed.

In pulse-modulated radar systems, pulses of r-f energy of about 1 to 50 μsec duration are radiated in a sharply directed beam. After the pulse has been emitted, the transmitter is shut off and a receiver, nearby, is turned on to receive the echo from a material object. When all echoes from one pulse have been received the transmitter is turned on again to send out another pulse, and the process is repeated at a rate of perhaps 1,000 times per sec. The echoes of pulses are picked up by a directive antenna and are displayed as a pattern on the screen of a cathode-ray tube. From the patterns produced by such echoes, the range and direction, or the distance and elevation of distant objects can be determined.

The directional antennas for transmission and reception in radar systems operate conjointly. This is automatically achieved if the same antenna is used for reception as for transmission. The pulse duration and repetition frequency of signals emitted by the transmitter must be synchronized and correlated with the scanning circuits of the cathode-ray-tube system in the receiver to produce a pattern that can be interpreted in terms of range and direction. Radar systems therefore represent a kind of radio-communication method in which both the wave directive properties and the demodulating and modulating operations of the transmitter and receiver must be simultaneously synchronized in a wholly coordinated system.

19. Radio System as a Whole. One purpose of this chapter is to outline, broadly, the general method of operation of various representative types of radio-communication systems. Another is to show how the complexity of the systems—and especially the interdependence of transmitter and receiver design—is related to the rate at which information is conveyed. The interdependence of receiver and transmitter design is especially important so far as concerns: (1) frequency of carrier wave; (2) the nature, timing, or synchronizing of the modulation and demodulation functions; and (3) the correlation or synchronization of the time and space directivity patterns of the emitting and intercepting apparatus.

20. Functional Elements. A block diagram representing the functional elements of a radio-communication system is shown in Fig. 1. For simplicity, the sources of power with which the equipment must be associated have been omitted. The information to be transmitted, and represented by $i_1(t)$, is converted into electrical form by a transducer. The output of the transducer, together with the output of a carrier generator, is fed to a modulator unit where the carrier wave is modulated by the signal wave in accordance with the information represented by $i_1(t)$. The modulated carrier wave $e_1(t)$ is radiated and propagated in the desired direction by an emitter, which may be a wave guide, transmission line, or antenna.

At the receiving end of the system, the intercepted energy is converted into voltage variations $e_2(t)$ by a wave interceptor or antenna. By amplifying, conversion, and demodulating apparatus, $e_2(t)$ is (usually) magnified and is converted into sense-perceptory energy by the transducer whose output is $i_2(t)$.

The demodulating and wave-intercepting operations at the receiver may require timing or synchronizing operations related to the modulating and wave-emitting operations at the transmitting end. Such timing and synchronizing operations may employ frequency-control apparatus at each end of the system, or the timing and synchronizing operations at the receiving end may be controlled by suitable timing pulses or signals generated in the transmitter, present in the transmitted wave, and

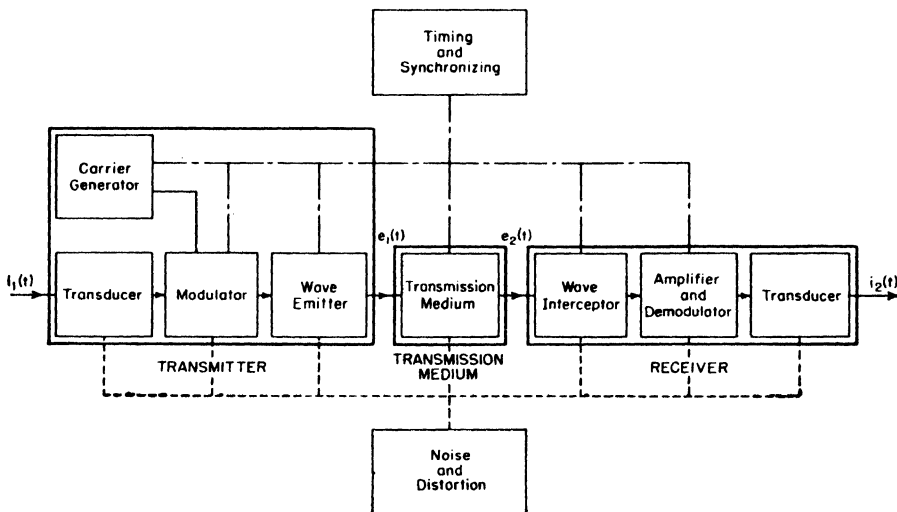



FIG. 1. Functional elements of complete radio-communication system. Signal or intelligence $i_1(t)$ is fed into the transmitter and emerges as electrical function of time $e_1(t)$, passes through transmission medium whose output is $e_2(t)$, and through receiver whose output is $i_2(t)$. Noise and distortion may be introduced at several points, as indicated by dashed lines. In certain systems timing or synchronizing operations between the transmitter, transmission system, and receiver may be required (as in television or radar systems, for example) in order for the system to function properly.

actuating or initiating the operations of frequency or time-control circuits at the receiving end. Such timing and synchronizing functions are indicated in Fig. 1 as being common to both ends of the system. Because such interlocking frequency-control, timing, or synchronizing methods are not required in all radio systems, they are indicated by dashed lines rather than by solid lines.

Noise and distortion may be introduced into the system at a number of different points, although usually most noise is introduced in the transmission system as shown by the dashed lines. Once introduced, however, the deleterious effects of noise and distortion are encountered at all points to the right of their introduction in Fig. 1.

21. Signal Frequencies vs. Amount of Information Transmitted. To be useful in representing intelligence, each symbol expressing a certain amount and kind of intelligence must be different from every other symbol expressing a different amount and kind of intelligence. In other words, each symbol, along with the meaning it expresses, must be unique. Using such symbolic notation, the communications engineer is interested in determining the mechanism by which a quantitative measure of nformation can be developed.

The amount and precision of information or intelligence conveyed depend upon the

number and sequence of the symbols selected. Hence the number of symbols required to convey the desired amount of information may be taken as a quantitative measure of information. The significance or meaning of the intelligence transmitted depends not alone on the number of symbols selected, but also upon their arrangement or the order or sequence of selection. This point of view has the great practical advantage of being objective and completely divorced from physiological factors.

If information is measured quantitatively by the number of selections required, the rate at which information is conveyed depends upon the number of symbols selected per unit of time, or upon the rate of speed of symbol selection. It follows that the rate of conveying intelligence depends upon the frequency with which elementary choices are made to select the symbols. The common and frequently used symbols may be represented or expressed by simple or easy choices or selections, whereas the unusual or infrequently used symbols may be represented or expressed by a more complex selection process requiring more choices for final selection.

A band of frequencies is required to transmit information, and the band needed in any particular case depends upon the rate at which information is transmitted. Typical values of signal band widths normally required for several different kinds of communication systems are given in Table 1.

The relationship between band width and information was first clearly stated by Hartley, who enunciated the following important principle of communication:[1]

The maximum rate at which information may be transmitted over a system whose transmission is limited to frequencies lying in a restricted range is proportional to the frequency range. From this it follows that the total amount of information which may be transmitted over such a system is proportional to the product of the frequency range which it transmits by the time during which it is available for transmission.

Table 1. Signal-frequency Band Widths Required for Various Kinds of Radio-communication Service

Kind of Service	Typical Band Widths, Cps
Manually keyed telegraphy (25 words/min)	0–100
Tone-modulated telegraphy (speed of 25 words/min, 1,000 cps tone)	0–2,100
Commercial telephony, single side band	0–3,000
Commercial telephony, double side band	0–6,000
High-quality single-side-band telephony	0–5,000
High-quality double-side-band telephony	0–10,000
Exceptionally high quality single-side-band telephony	0–20,000
Exceptionally high quality double-side-band telephony	0–40,000
Facsimile, 8- × 10-in. picture, 100 lines/in. transmitted in 10 min with 1,000-cps tone	0–4,000
Low-definition television (180 lines, 24 frames/sec)	24–410,000
Good-definition television (525 lines, 30 frames/sec)	30–4,350,000
High-definition television (1,029 lines, 30 frames/sec)	30–16,650,000
Frequency-shift telegraphy	0–1,800
Commercial frequency-modulated telephony	0–36,000
High-quality frequency-modulated telephone broadcasting	0–200,000
Frequency-modulated facsimile (8- × 10-in. picture, 100 lines/in., transmitted in 10 min with tone of 1,000 cps)	0–14,000

22. Modulation. The information to be transmitted is contained in, or represented by, the frequency spectrum of the signal or message. The message is adapted to transmission by converting signal variations into single-valued time functions of electrical energy through the use of suitable transducers. For all practical purposes in well-designed systems, the frequency spectrum of its electrical representation is the same as that of the original form of the message. Electrical energy at signal frequencies is not usually in a form useful for direct transmission through the medium of propagation, but it can be converted into proper form by a process of modulation in which the properties of the carrier wave and the signal wave are combined.

A carrier wave has such properties as will enable it to be propagated through the transmission medium. A signal wave is a wave conveying some intelligence, message, or effect. Modulation is the process of modifying some characteristic of a carrier wave

[1] Hartley, R. V. L., Transmission of Information, *Bell System Tech. J.*, **7** (3), 535, July, 1928.

in accordance with the signal wave, the result of such modification being a modulated wave, which is therefore a combination of the carrier and signal waves.

Information may be transmitted continuously, or it may be transmitted intermittently. The first method is represented by voice signals or music, the second by radar pulses or by the characters of a telegraph code. These two ways of transmitting information give rise to two fundamentally different methods of modulation—continuous, and intermittent or pulse.

In continuous methods of modulation, some property of the sinusoidal carrier frequency is continuously altered by the information to be transmitted. If the amplitude of the carrier is modified, amplitude modulation results. The argument of the sinusoidal carrier may also be modified by the signal. In this case angle rather than amplitude modulation results. Angle modulation may be subdivided into two related types. If the frequency of the carrier is varied by the signal, frequency modulation results, whereas phase modulation results if the phase of the carrier is altered.

In pulse modulation, the information is not projected continuously, but rather as discrete signal elements, usually short pulses. The message is sampled by the pulses, and some property of the pulses is then modified by the character of the signal to be transmitted. Pulse-modulation methods may be divided into unquantized or ordinary pulse modulation, and quantized or pulse-code modulation methods. By *quantization*, we mean that a large range of quantities may be divided into a set of small ranges, each of which is assigned a single discrete quantity. Quantization permits the approximate representation of a continuously varying function by a set of discrete values. The difference between two adjacent discrete values is called a quantum.

If the amplitude of the pulses is continuously variable in accordance with the signal to be transmitted and the timing of the pulses is unchanged or independent of the signal, ordinary or unquantized pulse-amplitude modulation results. If the time properties of the pulses are continuously altered, pulse-time modulation results. Thus, by modifying the duration, position, or frequency of the pulses continuously in accordance with the signal, ordinary or unquantized pulse-time modulation may be designated, respectively, as pulse-duration modulation, pulse-position modulation, or pulse-frequency modulation.

In pulse-code modulation the signal wave is sampled at discontinuous intervals determined by pulses, and is then quantized and coded. Quantization is accomplished by rounding off the value of the signal to the nearest one of a finite set of permitted values at each sampling step. Such quantization permits the signal to be transmitted by codes, such as the familiar telegraph code.

No matter whether continuous or pulse-modulation methods are involved, the modulated wave contains new frequencies not present in either the carrier or the signal wave. These new frequencies are composed of the sums and differences of the carrier and signal frequencies, and are called side-band frequencies. The signal frequencies are usually low compared with the carrier frequency so that, in most practical cases, either or both side bands are of substantially the same frequency as that of the carrier; it is the high-frequency characteristic of the modulated wave imparted to it by the carrier that permits it to pass through the propagation medium.

All essential signal information is contained in either one of the side bands. It is unnecessary, therefore, to transmit both side bands, or, for that matter, even the carrier frequency. In most radio systems, the carrier frequency is transmitted; otherwise it must be provided locally in the receiver. Total or partial elimination of one of the side bands reduces the frequency spectrum or channel band width required for transmitting a given kind of signal.

Since modulation is a complicated process, the question naturally arises as to why modulation is necessary. Most importantly, it imparts intelligence to the carrier. In radio communication, and also in many kinds of telegraphy and telephony, modulation is necessary as a means for conveying signals from one point to another. Modulation is also employed to achieve desirable characteristics in communication. For

example, the signal-to-noise ratio may be improved through suitable methods of modulation, the effects of natural or man-made static may be drastically reduced, or the frequency spectrum available for communication services may be more effectively utilized.

23. Frequency and Phase in Modulated Waves. A sinusoidal variation with constant amplitude and uniform phase shift per unit time is characterized by a fixed and constant single frequency. Such a constant frequency is a feature of an unmodulated carrier wave and is commonly encountered in communication systems. When the carrier is modulated, however, the resultant wave is no longer characterized by a single constant frequency or phase shift, and it is necessary to devise new concepts and definitions for the instantaneous values of frequency and phase of waves more complicated than simple harmonic variations.

Frequency is the time rate of change. Hence the angular frequency ω is the time rate of change of the phase angle θ; conversely phase angle θ is the integral, with respect to time, of the frequency ω. With these concepts and definitions, we are no longer required to deal with constant values of frequency and phase shift; instead we can define instantaneous values of these terms.

The instantaneous angular frequency, in radians per second, is

$$\omega_i = \frac{d}{dt}\,\theta(t)$$

when the phase shift is a function of time $\theta(t)$ and is equal to $2\pi f$, where f is the frequency in cps.

When the phase shift and frequency vary with time, the instantaneous value of the phase shift is

$$\theta(t) = \int \omega_i \, dt$$

where ω_i is the instantaneous angular frequency in radians per second.

The expressions for instantaneous frequency and instantaneous phase shift are applicable, of course, to the case when these factors are constant with respect to time. But the particular value of these terms lies in their application to time-varying phenomena, such as are encountered in processes of modulation.

24. Amplitude Modulation. In amplitude modulation the amplitude of the carrier wave is varied in accordance with the signal. The resultant modulated wave is no longer of constant single frequency but contains frequencies that are the sums and differences of the carrier and signal frequencies. The new frequencies are called side frequencies; if the signal is composed of a single sinusoidal wave, the spectrum of the modulated wave will contain the carrier frequency together with two side frequencies, one above and one below the carrier frequency and displaced from it, as mirror images, by an amount equal to the signal frequency. If the signal is composed of a line or continuous band spectrum, the side bands will also be line or continuous band spectra as mirror images on both sides of the carrier frequency. The side bands contain the information imparted to the carrier wave by the signal in the process of modulation. If allowance is made for the fact that the modulation process shifts the frequency of the signal by an amount equal to the carrier frequency, the spectrum of each of the side bands is usually the same as that of the original modulating wave or signal.

Consider a sinusoidal carrier wave of instantaneous value

$$i = I_c \cos(\omega_c t + \phi_c)$$

where I_c = amplitude of the carrier wave

$\omega_c = 2\pi f$ = angular frequency of carrier wave, radians per sec, and f = cyclic frequency, cps

ϕ_c = relative phase of carrier wave at time $t = 0$

Let the carrier be amplitude-modulated by $f(t)$, a sinusoidal time variation with angular frequency ω_s, of relative phase ϕ_s. Let

$$f(t) = [1 + m \cos(\omega_s t + \phi_s)]$$

where m is the modulation factor and lies between 0 and 1. Such a signal wave always has positive value and when applied to vary the amplitude of the carrier wave produces a modulated wave expressed by

$$\begin{aligned} i &= I_c[f(t)] \cos(\omega_c t + \phi_c) \\ &= I_c \cos(\omega_c t + \phi_c) && \text{carrier} \\ &\quad + \tfrac{1}{2} m I_c \cos[(\omega_c + \omega_s)t + (\phi_c + \phi_s)] && \text{upper side band} \\ &\quad + \tfrac{1}{2} m I_c \cos[(\omega_c - \omega_s)t + (\phi_c - \phi_s)] && \text{lower side band} \end{aligned}$$

In the process of modulation, the constant amplitude of the carrier has been modified by the signal wave $f(t)$, which now serves to define or outline the envelope of the modulated carrier wave. Since the amplitude is no longer constant, the single-frequency carrier has been expanded into a line spectrum, and, in general, at any instant, the modulated wave has an "instantaneous frequency" different from that of the carrier frequency.

The communication band width is the frequency spectrum occupied by the carrier and all upper and lower side-band frequencies. The carrier frequency is a single frequency conveying no information, but the above equations show that either the upper or the lower side band alone contains all the information of the original signal, by including the terms in ω_s. The frequency of the side bands has been shifted, however, with respect to the signal frequency, by an amount equal to the carrier frequency.

In general, the relative phase ϕ_s of each component of signal frequency in the modulated wave must be the same as that in the original form of the signal. This requirement may be relatively unimportant in some forms of communication, as in speech and music, since the ear is not sensitive to changes in phase shift; on the other hand, it is sometimes very critical, as in transmission of video signals for television.

If the modulation index m is zero, no signal information is conveyed to the carrier, whereas if $m = 1$, the envelope of the carrier wave (for sinusoidal signal wave) varies from zero to twice the value of its unmodulated amplitude. The average power in the carrier (which is proportional to I^2) remains constant. The power in each of the side bands cannot exceed 25 per cent of the average power in the carrier, and this occurs when $m = 1$. The total power in both side bands, therefore, cannot exceed half of that of the carrier for normal transmission of the modulated carrier. If the amplitude of the signal is increased relative to that of the carrier so that m exceeds unity, the carrier is said to be overmodulated; distortion results, and the useful power output in the side bands decreases.

Although the oldest and simplest form of modulation, amplitude modulation is not particularly effective or efficient compared with other methods of modulation.

25. Angle Modulation. In angle modulation the angle of a sinusoidal carrier is the parameter subject to variation by the modulating wave. Two forms of angle modulation are of practical value, (1) frequency modulation and (2) phase modulation. These forms of modulation are related, and sometimes both exist simultaneously in the same system.

In *frequency modulation* the instantaneous frequency of a sinusoidal carrier is made to depart from the constant unmodulated carrier frequency by an amount that is proportional to the instantaneous value of the signal or modulating wave.

To derive an expression for the current of a f-m wave, let the carrier current be

$$i = I_c \cos(\omega_c t + \phi_c)$$

and let the signal or modulating wave be $-\Delta\omega_c \sin \omega_s t$. Then, the instantaneous frequency of the modulated wave is

$$\omega_i = (\omega_c - \Delta\omega_c \sin \omega_s t) = \frac{d}{dt}\,\theta(t)$$

Now, $\theta(t) = \int \omega_i\,dt = \omega_c t + (\Delta\omega_c/\omega_s) \cos \omega_s t$, so that the modulated wave takes the form

$$i = I_c \cos\left(\omega_c t + \frac{\Delta\omega_c}{\omega_s} \cos \omega_s t\right)$$

For sinusoidal modulating waves, the frequency deviation of a f-m wave is $\Delta\omega_c/\omega_s$ radians per sec.

If the above expression is expanded and the terms grouped according to frequency components,

$$i = I_c \sum_{n=-\infty}^{\infty} J_n\left(\frac{\Delta\omega_c}{\omega_s}\right) \cos\left[(\omega_c + n\omega_s)t + \frac{n\pi}{2}\right]$$

where $J_n(\Delta\omega_c/\omega_s)$ is a Bessel function of the first kind and nth order with argument $\Delta\omega_c/\omega_s$.

The degree of modulation is usually defined as the ratio of the actual frequency deviation $\Delta\omega_c$ to the maximum frequency deviation $\Delta\omega_{c_{max}}$ of which the system is capable, or which is permitted by law.

In *phase modulation* the linearly increasing phase of the unmodulated carrier has added to it a time-varying phase angle whose magnitude is proportional to the amplitude of the applied modulating wave or signal.

To write an expression for the phase-modulated wave, the time-varying component of the phase angle must be proportional to the amplitude of the applied signal wave. Let the signal wave be $\Delta\phi \cos \omega_s t$ and let the unmodulated carrier be $i = I_c \cos (\omega_c t + \phi_c)$ as before. Then the modulated wave becomes

$$i = I_c \cos [\omega_c t + (\phi_c + \Delta\phi \cos \omega_s t)] = I_c \cos \phi(t)$$

where $\Delta\phi$ is the phase deviation factor of the resulting wave, expressed in radians. The instantaneous radian frequency ω_i is the time rate of change of phase angle $\theta(t)$ or

$$\omega_i = \omega_c - \omega_s \, \Delta\phi \sin \omega_s t$$

For sinusoidal modulating wave, the frequency deviation of the phase-modulated wave is $\omega_s \, \Delta\phi$ radians per sec.

If the above expression for the modulated wave is expanded, and terms are grouped according to frequency components, the phase-modulated wave is expressible as

$$i = I_c \sum_{n=-\infty}^{\infty} J_n(\Delta\phi) \cos\left[(\omega_c + n\omega_s)t + \frac{n\pi}{2}\right]$$

where $J_n(\Delta\phi)$ is a Bessel function of the first kind and nth order with argument $\Delta\phi$.

In *phase modulation*, the phase $\theta(t) = \phi_c + \Delta\phi \cos \omega_s t$ is varied in accordance with the signal. The quantity $\Delta\phi \cos \omega_s t$ is called the phase deviation and is expressed in radians. The degree of modulation is usually defined as the ratio of the actual phase deviation $\Delta\phi$ to the maximum phase deviation $\Delta\phi_{max}$ of which the apparatus is capable at any moment, or is permitted by law. Thus, for phase modulation the degree of modulation is defined in terms of the properties of the system in which it is used, and not alone in terms of properties of the signal and carrier, as in the case of amplitude modulation.

26. Some Properties of Angle-modulation Systems. In f-m systems the deviation of the instantaneous frequency from its unmodulated value is proportional to the magnitude of the modulating wave and is independent of the frequency of the modulating signal. On the other hand, in phase-modulation systems the deviation of the instantaneous frequency from its unmodulated value is proportional to the derivative of the modulating wave. Thus, although phase modulation and frequency modulation are particular forms of angle modulation, they are not essentially different, and one can be converted into the other.

There is, in fact, no important difference between a frequency- and a phase-modulated wave so long as modulation is accomplished by a single sinusoidal frequency, except a difference in degree of modulation. If the modulating signal has components of many

frequencies, however, then there is definitely a difference; f-m phase-modulated signals have greater deviation at the higher modulating or signal frequencies.

Angle-modulation systems have important noise-reduction properties, particularly for random noise, provided the signal level exceeds the noise level. When the signal-to-noise ratio is greater than unity, and preferably is greater than unity with a safe margin, the noise-reduction properties of angle-modulation methods are primarily due to two factors. The modulating signal produces a much larger deviation of the carrier frequency than is produced by a random distribution of side bands. In addition, superposition of random noise on a wide-band angle-modulated carrier produces modulation frequency components too high to pass through an audio system, so that these components are not heard as noise. Only the side bands produced by noise whose frequency components are close to that of the unmodulated carrier produce noise in f-m systems, and these are relatively ineffective in producing f.m. The signal has effective side bands far removed from the carrier, and all of these act together to produce the swing or deviation at the signal frequency.

27. Carrier and Side-band Suppression. When a carrier of constant angular frequency ω is modulated by a signal, the frequency of the modulated carrier is broadened. The modulated carrier contains frequencies of the carrier frequency, and bands of frequencies above and below the carrier frequency and displaced from it by the frequency of the signal component. These frequencies above and below the carrier frequency are called the upper and lower side-band frequencies, respectively. The frequency spectrum occupied by the carrier and both side bands is called the communication band width.

In amplitude modulation, the communication band width is twice the frequency of the highest-frequency component of the original signal which is transmitted. Thus a broadcast channel having a band width of 10 kc and transmitting both side bands is capable of transmitting signal components in each side band up to 5 kc; to provide some spectrum separation between adjacent channels, signals of somewhat less than 5 kc could be accommodated in channels 10 kc wide.

In angle modulation, there exists an infinite number of pairs of harmonic side frequencies, spaced above and below the carrier by positive multiples of a single signal frequency. Not all these frequencies are important, however, and the maximum value of n which need be considered for a given modulation index m is expressed empirically by the relation

$$n_m = m + km^{0.27}$$

where $k = 2.4$ for $J_n(m) = 0.01$, and $k = 3.5$ for $J_n(m) = 0.001$.

Since all the information of the signal is contained in either of the side bands, it is unnecessary to transmit both side bands. One side band may be removed, and the portion of the frequency spectrum from which the side band is removed will then be available for other services. One of the side bands may be removed by means of filters having very sharp cutoff properties, so that the carrier and one side band are transmitted while the other side band is eliminated. Single-side-band transmission may also be accomplished by removing one of the side bands by means of a phase-balancing process instead of filtering. In the phasing method, two carriers are generated with 90-deg phase shift between them, and these are modulated by two signals which also have a 90-deg phase shift between them. The output of the two modulators consists of the carrier and one side band, the other side band being removed by cancellation of phases in the two modulating units.

Sometimes it is not desirable or feasible to remove all of one side band, but merely most of it, particularly at the higher signal-frequency components. *Vestigial side-band transmission* then results.

Carrier suppression is the process of removing the carrier component in a modulated signal, leaving one or both of the side bands. Removal of the carrier results in no loss of signal for the carrier contains no information. An appreciable saving of transmitted power may be achieved by not transmitting the carrier, with resultant increase

in the transmitter efficiency, for in the absence of a modulating signal no power is transmitted. Even when signals are transmitted, a larger fraction of the transmitter power may be applied to the side bands instead of delivering it to a carrier conveying no information. For carrier-suppression systems, a carrier of proper frequency and phase must be supplied at the receiver. Since it is not always easy to achieve local generation of a carrier of proper frequency and phase, some systems transmit a portion of the carrier. The low-amplitude carrier transmitted in this case may then be used to control the locally generated carrier in the receiver.

28. Need for High-frequency Carriers. For intelligence-bearing signals to be propagated by electromagnetic waves, it is necessary that the electric and magnetic components of the wave interact with each other to support and sustain the process of propagation. This condition requires the generation of displacement currents of appreciable magnitude, and these are produced when the carrier frequency is comparatively high. Practical radio communication depends upon the employment of carrier frequencies greater than (and usually very much greater than) about 10 kc; carrier frequencies may extend to several thousand megacycles, *i.e.*, to the region of 10^9 cps or more.

The band width of the modulated carrier should be a small fraction of the unmodulated mean carrier frequency. This limitation requires that the carrier frequencies be very much greater than the highest signal frequency to be transmitted. This restriction also requires very high carrier frequencies for those kinds of communication in which a considerable amount of information must be transmitted in a short interval of time (as in television or radar systems) or where the properties of the system necessitate reasonably large excursions from the mean carrier frequency (as in f-m systems).

Table 2. Relations between Carrier Frequency and Band Widths for Various Classes of Radio Communication

Type of service	Typical carrier frequency, Mc	Typical band width, cps	Ratio of max band width to carrier frequency
Long-wave telegraphy	0.100	0–100	0.001
Aeronautical telephony	0.200	100–3,000	0.015
Mobile (emergency) telephony	0.50	100–3 000	0.006
Broadcast telephony	0.600	40–5,000	0.0083
Broadcast telephony	1.00	40–5,000	0.005
Broadcast telephony	1.50	40–5,000	0.003
Loran	1.90		
Standard frequency (tone modulation)	2.50	600	0.00024
Amateur telephony	7.25	100–3,000	0.00041
Emergency frequency	8.28		
Short-wave broadcasting	11.50	40–5,000	0.0004
Television channel 2	54–60	30–4,500,000	0.082
Marker beacon	75		
Television channel 6	82–88	30–4,500,000	0.055
F-m broadcasting	88	20–20,000	0.00024
F-m broadcasting	108	20–20,000	0.00018
Emergency frequency	140	100–3,000	0.00002
Television channel 7	180–186	30–4,500,000	0.025
Television channel 13	210–216	30–4,500,000	0.021
Long-wave radar (300 cm)	100	0–200,000	0.002
Microwave radar (10 cm)	3,000	0–500,000	0.0001
Microwave radar (3 cm)	10,000	0–500,000	0.00005

The ratios of the maximum signal frequency to the (mean) carrier frequency for typical classes of radio service are given in Table 2. These data show that the band width for the modulated carrier frequency is, at most, but a few per cent of the mean carrier frequency for practical systems.

29. Method of Solving Electrical-communication Problems. Solving many problems in electrical communication involves the following steps:

1. The differential equations expressing the dynamic equilibrium of the system are set up with initial and boundary conditions specified.
2. The functions satisfying those sets of equations are determined by a process of integration which yields the formal solution to the problem. The formal solution thus found is more general than is required for a particular solution.
3. The formal solution must then be subjected to the initial conditions of the system, as well as to the boundary conditions of time and space to obtain the solution for the particular problem under consideration.

The need for analytical treatments of communication problems thus becomes clear. To be useful for communication purposes, however, the differential equations of dynamic equilibrium of the communication system, as well as their formal solutions, must be expressible in terms of information-bearing signals. It becomes necessary, therefore, to obtain quantitative expressions for the representation of intelligence in developing an analytical approach to communication engineering. Such expressions should be as simple as possible, for mathematical manipulation, consistent with the requirement that they be sufficiently rigorous and useful for engineering and scientific work.

References

Principles of Electrical Communication:

Armstrong, E. H.: A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation, *Proc. IRE*, **24** (5), 689, May, 1936.
Black, H. S.: "Modulation Theory," Van Nostrand, 1953.
Bureau of Ships, Navy Department: "Radar System Fundamentals," NAVSHIPS 900,017, Washington, D.C., 1944.
Carson, J. R.: Notes on the Theory of Modulation, *Proc. IRE*, **10** (1), 57, February, 1922.
Cherry, E. Colin: A History of the Theory of Communication, *Proc. AIEE*, **98**, Part III (55), September, 1951.
Colpitts, E. H., and O. B. Blackwell: Carrier Current Telephony and Telegraphy, *Trans. AIEE*, **40**, 205, Feb. 17, 1921.
Deloraine, E. M.: Pulse Modulation, *Proc. IRE*, **37** (6), 702, June, 1949.
Fink, D. G.: "Television Engineering," 2d ed., McGraw-Hill, 1952.
———: "Television Engineering Handbook," McGraw-Hill, 1957.
Gabor, D.: New Possibilities in Speech Transmission, *J. Inst. Elec. Engrs. (London)*, **94**, Part III, 369, 1947; **95**, Part III, 39, 412, 1948.
———: Theory of Communication, *J. Inst. Elec. Engrs. (London)*, **93**, Part III (26), 429, November, 1946.
Hartley, R. V. L.: Transmission of Information, *Bell System Tech. J.*, **7** (3), 535, July, 1928.
Hershberger, W. D.: "Principles of Communication Systems," Prentice-Hall, 1955.
Metz, P., and F. Gray: A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television, *Bell System Tech. J.*, **13** (3), 464, July, 1934.
Nyquist, H.: Certain Factors Affecting Telegraph Speed, *Bell System Tech. J.*, **3**, 324, 1924.
———: Regeneration Theory, *Bell System Tech. J.*, **11** (1), 126, January, 1932.
———: Thermal Agitation of Electric Charge in Conductors, *Phys. Rev.*, **32**, 110, July, 1928.
Osborne, H. S.: Principles of Electric Circuits Applied to Communication, *Bell System Tech. J.*, **8** (1), 3, January, 1929.
Reintjes, J. F., and G. T. Coate: "Principles of Radar," 3d ed., McGraw-Hill, 1952.
Rice, S. O.: Mathematical Analysis of Random Noise, *Bell System Tech. J.*, **23** (3), 282, July, 1944; **24** (1), 46, January, 1945.
Shannon, C. E.: A Mathematical Theory of Communication, *Bell System Tech. J.*, **27** (3), 379, July, 1948; **27** (4), 623, October, 1948.
——— and W. Weaver: "The Mathematical Theory of Communication," University of Illinois Press, 1949.
Tuller, W. G.: Theoretical Limits on the Rate of Transmission of Information, *Proc. IRE*, **37** (5), 468, May, 1949.
Van der Pol, B.: Frequency Modulation, *Proc. IRE*, **18** (7), 1194, July, 1930.
Wiener, N.: "Cybernetics," Wiley, 1948.

Mathematical Principles:

Byerly, W. E.: "Fourier Series and Spherical Harmonics," Ginn, 1893.
Carson, J. R.: "Electric Circuit Theory and the Operational Calculus," McGraw-Hill, 1926.
Franklin, P.: "Differential Equations for Electrical Engineers," Wiley, 1934.
Guillemin, E. A.: "The Mathematics of Circuit Analysis," Wiley, 1949.
Kármán, T. V., and M. A. Biot: "Mathematical Methods in Engineering," McGraw-Hill, 1940.
Shannon, C. E., and W. Weaver: "The Mathematical Theory of Communication," University of Illinois Press, 1949.
Wiener, N.: Generalized Harmonic Analysis, *Acta Math.*, **55**, 118, 1930.

———: "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series," Wiley, 1949
———: "The Fourier Integral and Certain of Its Applications," Cambridge, 1933.

Communication Systems:

Bureau of Ships, Navy Department: "Radar System Fundamentals," NAVSHIPS 900,017, Washington, D.C., 1944.
Fink, D. G.: "Television Engineering," 2d ed., McGraw-Hill, 1952.
Goldman, S.: "Frequency Analysis, Modulation, and Noise," McGraw-Hill, 1948.
Goodall, W. H.: Telephony by Pulse Code Modulation, *Bell System Tech. J.*, **26** (3), 395, July, 1947.
Hershberger, W. D.: "Principles of Communication Systems," Prentice-Hall, 1955.
Oliver, B. M., J. R. Pierce, and C. E. Shannon: The Philosophy of P.C.M., *Proc. IRE*, **36** (11), 1324, November, 1948.
"Radio Spectrum Conservation," Report of the Joint Technical Advisory Committee, IRE-RTMA, McGraw-Hill, 1952.
Reintjes, J. F., and G. T. Coate: "Principles of Radar," 3d ed., McGraw-Hill, 1952.

Information and Communication Theory:

Black, H. S.: "Modulation Theory," Van Nostrand, 1953.
Fry, T. C.: "Probability and Its Engineering Uses," Van Nostrand, 1928.
Goldman, S.: "Frequency, Analysis, Modulation, and Noise," McGraw-Hill, 1948.
Hershberger, W. D.: "Principles of Communication Systems," Prentice-Hall, 1955.
Uspensky, J. V.: "Introduction to Mathematical Probability," McGraw-Hill, 1937.
Wald, A.: "Sequential Analysis," Wiley, 1947.
Wiener, N.: "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series," Wiley, 1949.
———: "Cybernetics," Wiley, 1948.

ANALYTICAL APPROACH TO COMMUNICATION ENGINEERING

30. Theories of Communication. The general problem of electrical communication is to: (1) accept, at the transmitting end of the system, the message to be conveyed, (2) transform the message into such form as will permit its passage through the communication system, (3) convey or transmit the electrical form of the message from a transmitting point to one or more receiving points, and (4) re-create a satisfactory replica of the original message at the receiving end of the system.

The term "message" is to be interpreted in a broad sense; it may mean any information-bearing function, *i.e.*, any concept or entity translatable into physical form capable of being measured or specified as a single-valued time function. The character of the message determines important features of the system (such as band width and power requirements) especially when considered in relation to interference such as distortion, noise, cross talk, and the like. As used in the above sense, the term "message" is synonymous with the term "information."

Views as to what constitutes information, and the means by which information can be expressed analytically, have had important bearings on the theories underlying communication engineering and the development of communication systems. Two theories have been effective in building up our present understanding of systems of communication, control, and information processing. These are the classical theory, which dominated thinking until the Second World War, and the statistical theory developed since about 1942.

31. Classical Theory. The classical theory of communication depends primarily upon application of the Fourier series and Fourier integral methods of expressing the single-valued time functions of the information to be transmitted. The Fourier series makes use of a number (which may be infinite) of harmonic functions of time to represent single-valued magnitude-time functions under steady-state conditions. The use of Fourier series implies that the harmonic functions representing the wave form have existed for an infinite length of time, and assumes that they will continue, unchanged, for an infinite period. The Fourier integral represents a single-valued aperiodic or transient magnitude-time function whose components, again, are assumed to exist unchanged for all time.

The wave form specified by a Fourier series may be expressed as a magnitude-time function in the time domain, or as a pair of line spectra in the frequency domain; in the latter case the amplitude and phase of each discrete harmonic component are

plotted as line functions of the frequency. A wave form expressed by a Fourier integral may likewise be represented in the time domain as a magnitude-time function; in the frequency domain it is represented by a pair of continuous amplitude and phase functions of frequency. A given function expressed in the time domain can be transformed into the frequency domain and vice versa by means of Fourier transformations. Fourier transforms have played an important role in classical communication theory. The classical point of view has been useful in determining important properties of the communication system, so long as the signals to be transmitted can be represented by periodic or aperiodic time functions.

According to the classical point of view, the recurrent or transient functions expressible by the Fourier series or integrals are the commodity which the communication system handles; such functions represent the information to be conveyed. If the information is to be expressed in terms of recurrent or transient wave forms, the Fourier theory determines the values of the information function for all values of time—past, present, and future. If a message is represented by a function of the type whose values are specified once and for all, for all values of the independent variable (usually time) the information expressed in this way must be such that its entire nature is precisely known and predictable. This point of view is in direct conflict with the very nature of information, for, to be useful, the flow of information must contribute some knowledge not previously known, and consequently must be unpredictable.

32. Statistical Theory. The proper functioning of the communication system requires that the fluctuating phenomena, containing the information for transmission from one portion of the system to another, must ultimately be re-created at the receiving end. The fluctuations of information may be represented by an array of unpredictable but measurable quantities distributed in time. It is the unpredictable nature of the message which makes it carry information. The fluctuating phenomena representing the information are best described in terms of statistics and probability; these branches of mathematics form the backbone of modern information theory, rather than the harmonic or singularity functions on which classical communication theory was built. According to modern theory, messages and noise are random processes and are, accordingly, dealt with analytically as such.

Whatever their source, messages to be transmitted are regarded as random time-varying functions or time series. The statistical character of the message is assumed to remain constant with respect to time, and hence may be represented by time series. Time series are discrete or continuous sequences of quantitative data assigned to specific moments of time. They are studied and evaluated with respect to the statistics of their distribution in time. In the case of simple time series, a single numerical observation exists at each moment of the discrete or continuous sequence. Multiple time series consist of a number of separate quantities evaluated according to a time common to all.

The random functions used to represent information in modern communication theory can never specify definitely the value of phenomena at a specified instant of time in the future. Because they do have known or assumed statistical properties, however, certain matters regarding the *probability* of the existence of the phenomena can be ascertained. For example, it is possible to determine the probability that one letter in the English language will be followed by some other letter, or the probability that a picture element of given reflection density in a scanning process will be followed by another element of half that reflection density. That is, the structure of the information is not only determinable on a statistical basis, but it is also precisely determinable. In studying the statistical structure of messages by random time series, a number of statistical parameters will be useful. The correlation function, amplitude distributions, and certain conditional probability distributions are among the most important of these.

The mathematics of modern information theory, as well as the mathematics of classical communication theory, represents general relationships between significant entities in the communication process. As such, mathematics of communication

theory represents a method or technique of thinking rather than a set of formulas which may be used, in an automatic crank-turning process, to extract numerical results for specific design purposes. In this respect, at least, the mathematics of information theory is analogous to Maxwell's equations, for example, which, although they provide the underlying foundations from which circuit-design formulas can be derived, are not immediately amenable and directly applicable to the solution of circuit behavior.

33. Differences between Classical and Statistical Communication Theories. The fundamental difference between these theories of communication lies in the recently developed concept of the nature of information. The unpredictable property of information, and the use of the mathematics of statistics and probability theory to represent random processes, require knowledge of topics (such as generalized harmonic analysis and theory of statistical operations) which have been but recently introduced into the field of practical communication engineering.

The new point of view regarding communication theory complements and expands the older classical point of view; it does not entirely replace or invalidate it. In evaluating the performance of a communication system, there is still need to consider periodic and aperiodic functions. But the introduction of the statistical theory opens up new avenues which, at present, appear to give much promise for new technical achievements. The new approach seems to be most promising in pointing the way for conserving band width, or of trading band width for power (or vice versa), in the design of prediction circuits and filters for separating noise from the desired signal, and in comparing or evaluating a given system with its theoretical limit of performance.

34. Time Functions Encountered in Communication Engineering. In a communication system the measurable varying quantities that are distributed in time to represent information are operated upon or processed in their passage through the system in a variety of ways. The time functions or time series representing information are usually continuous—or may be treated as being continuous, without appreciable error—and may be classified as periodic functions, aperiodic functions, and random functions.

A *periodic function* is one whose amplitude repeats a set of values (usually, but not necessarily, continuous) after the passage of a given interval of time. Any single-valued recurrent wave form represents a periodic function. Periodic functions are important in communication systems; for example, the carrier waves used in radio and wire telegraphy and telephony are periodic functions. Any single-valued periodic wave can be completely specified by the wave form for a single cycle, or by its amplitude and phase spectra. Since recurrent cycles of periodic phenomena are alike, the specifications for one complete wave or cycle are the same as those for any other complete cycle. Because all successive cycles of periodic phenomena are alike and can be completely specified by the amplitudes and phases of the harmonic components as functions of frequency, or by constructing the wave form (both of which are derivable from a single cycle of recurrent phenomena), no new information is conveyed if more than one cycle of phenomena is received. Hence, periodic phenomena are incapable of representing a continuous flow of information; in fact the information they can convey is exceedingly limited.

An *aperiodic function* is one which is devoid of periodicity, or for which the period is recurrent only if infinitely long. Aperiodic functions of time are singularity functions and are usually associated with transient phenomena. For purposes of analysis, aperiodic functions are usually expressed in terms of the Fourier integral which specifies the value of the function for all values of time. Aperiodic functions may be expressed by the wave form of the function, or by the amplitude and phase spectra which are now continuous spectra, rather than line spectra as is the case for periodic phenomena. Again, it is possible to specify the wave form of an aperiodic phenomenon for all values of time, and for this reason aperiodic phenomena, like periodic phenomena, are incapable of representing a continuous flow of information.

Finally, *random functions* are encountered in communication systems. By a random function is meant a single-valued function in which the instantaneous values of

the fluctuating phenomena are not subject to precise prediction, but whose statistical properties can be precisely evaluated. It is impossible to specify accurately the magnitude of a random function at any instant of time; all we can do is determine the statistical probability that the function will have a designated value—or will lie within a specified range of values—at any given instant. The fact that the value of random functions cannot be accurately predicted for any specified instant is precisely that characteristic property which makes them so useful in representing the flow of information. For purposes of analysis, random functions are evaluated by applying principles of generalized harmonic analysis to the random or statistically determined functions.

The three types of functions encountered in communication systems—periodic, aperiodic, and random—may each be represented by means of generalized harmonic analysis. Harmonic analysis applied to periodic functions gives rise to the simplest mathematical forms, whereas harmonic analysis applied to random or statistically determined functions is usually the least familiar to, if not the most difficult for, most practicing communications engineers.

35. Periodic Functions. The simplest type of periodic function is that which varies harmonically with respect to time. Such a function may be expressed in terms of three constants by an equation of the form

$$y = f(t) = E \sin (\omega t + \theta) = E \cos (\omega t + \phi)$$

where E is the amplitude or maximum value of the function, $\omega = 2\pi f$ is the angular frequency in radians per second if f is the cyclic frequency in cps, and θ and ϕ are phase angles relating the zero value of the sine or cosine function, respectively, with that instant of time from which it is assumed that the function has its initial, zero, or starting value.

Harmonic functions are mathematically simple to manipulate and represent quite accurately wave forms frequently encountered in electrical-communication systems. Moreover, sinusoids provide the basis for dealing with much more complicated wave forms, expressible by means of the Fourier series, since any periodic wave form may be represented by, analyzed into, or synthesized from, a series of harmonic functions with properly selected values of the three constants for each component term.

FIG. 2. Harmonic function plotted explicitly against time.

Harmonic functions involve three constants: the amplitude E, the angular frequency ω (or the cyclic frequency f), and the phase displacement, given above by θ or ϕ. When these three constants are known the harmonic function (either the sine or the cosine) is completely specified. In the time domain, harmonic functions may be illustrated graphically by the familiar sine-wave curve of Fig. 2. Such a representation has the value of illustrating, pictorially, the manner in which the function varies with time. From such an amplitude-time curve, the three constants (amplitude, frequency, and phase) may be determined.

An amplitude-time curve, such as that shown in Fig. 2, is a satisfactory method of representing the manner in which the magnitude of a single time series varies with time. But all harmonic functions (also called *sinusoids*) have similar wave shapes, and once the manner in which the magnitude of a harmonic function varies with time is known, any sinusoid is just like any other, except for its three unique and characteristic constants; hence, it may be just as well—and more compactly—expressed in terms of its constants. Moreover, the magnitude-time wave is not a very satisfactory method of representing many—or even several—wave forms which may exist simultaneously in a circuit or system.

36. Amplitude- and Phase-frequency Spectra. In the frequency domain, the simple harmonic function may also be represented by means of its three characteristic constants on a graphical plot in which frequency is the independent variable, and amplitude is the dependent variable on one plot, and phase is the dependent variable on the other. The amplitude-frequency and phase-frequency plots (called a pair of spectra) for a simple harmonic function are shown in Fig. 3. The pair of spectra is capable of conveying all the information for a given harmonic function, like that contained in the time-domain plot of Fig. 2, and can represent much more compactly and systematically the necessary data for each separate harmonic function when many harmonic terms may be required to represent any periodic wave form more complicated than sinusoids.

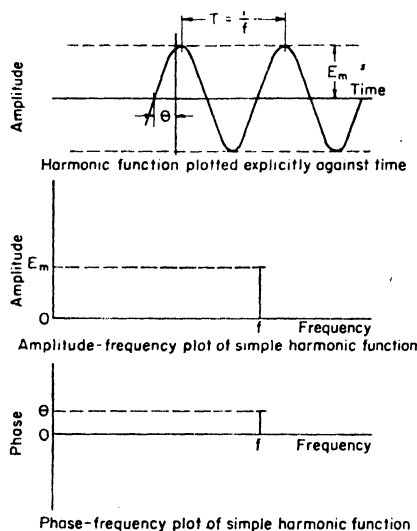



FIG. 3. Harmonic function plotted against time with its corresponding amplitude- and phase-frequency spectra plots. The spectrum plots convey the same information as does the plot in the time domain.

The middle diagram of Fig. 3 is called the *amplitude-frequency spectrum*, whereas the lower diagram is called the *phase-frequency spectrum*. Since the amplitude, frequency, and phase are all constant for a given sinusoid, the amplitude and phase spectra are represented merely as lines of appropriate length for the specified frequency. The line-spectrum form is especially useful for specifying the characteristics of each harmonic component, particularly when many components are needed to represent a complicated periodic wave. See also Fig. 7.

37. Analytical Expressions for Simple Harmonic Functions. Simple harmonic functions may be regarded as being generated by the projection, on a suitable axis, of a directed quantity (frequently called a *vector*, although the term *phasor* appears to be more appropriate in this application) of constant magnitude E_m, rotating with constant angular velocity ω radians per sec in a positive (counterclockwise) direction, with an initial phase displacement ϕ, as shown in Fig. 4. The projection of one end of the rotating phasor on the x, or real, axis generates the cosine function, whereas the projection of the tip of the phasor on the y, or quadrature, axis (usually designated as the *imaginary* or j axis) generates the sine function.

At any instant, the sine and cosine functions are the rectangular components of the rotating phasor, as expressed by the vector sum

$$E = E_m \cos (\omega t + \phi) + jE \sin (\omega t - \phi)$$

where $j = (-1)^{1/2}$ is an operator having the algebraic value indicated, and geometrically represents rotation of 90 deg (or $\frac{1}{2}\pi$ radians) in the positive or counterclockwise direction. The phasor may also be expressed in terms of its polar components—the amplitude and phase—at any instant, rather than in terms of its rectangular coordinates at the same instant. In polar coordinates,

$$E = |E_m|\underline{/\phi} = (E_{\mathrm{Re}}^2 + E_{\mathrm{Qu}}^2)^{1/2} \underline{/ \tan^{-1} \frac{E_{\mathrm{Re}}}{E_{\mathrm{Qu}}}}$$

In this mode of expression, $|E_m| = (E_{\mathrm{Re}}^2 + E_{\mathrm{Qu}}^2)^{1/2}$ is the amplitude or maximum value of the rotating phasor, expressed in terms of its rectangular coordinates $E_{\mathrm{Re}} = E_m \cos (\omega t + \phi)$ and $E_{\mathrm{Qu}} = E_m \sin (\omega t - \phi)$, whereas the phase angle is also expressed in terms of the rectangular orthogonal components, through the relation $\phi = \tan^{-1} (E_{\mathrm{Re}}/E_{\mathrm{Qu}})$.

Expressions in rectangular form are more convenient when harmonic functions are to be added or subtracted, whereas the polar form is more convenient when sinusoids are to be multiplied, divided, or raised to a power.

38. Exponential Representation of Harmonic Functions. It is shown in elementary textbooks on trigonometry (see, for example, C. I. Palmer, C. W. Leigh, and S. Kimball,

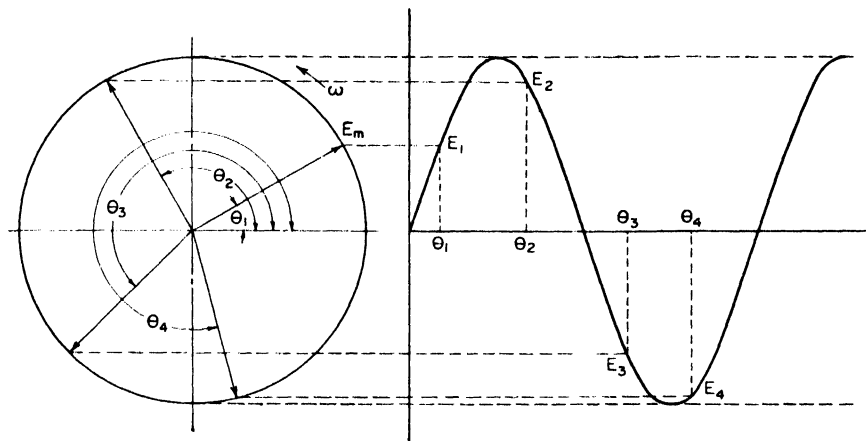



FIG. 4. Generation of harmonic function by means of phasor of magnitude E_m rotating in positive (counterclockwise) direction with constant angular velocity ω. Left, the path of the rotating phasor for four different angular displacements. Right, the harmonic function as generated by projection of the tip of the phasor for different times.

"Plane and Spherical Trigonometry," 5th ed., Chap. XI, McGraw-Hill, 1950) that the expansion for an exponential quantity in complex form is

$$e^{jx} = 1 \pm \frac{(jx)}{1!} \pm \frac{(jx)^2}{2!} \pm \frac{(jx)^3}{3!} \pm \cdots \pm \frac{(jx)^n}{n!} \pm \cdots$$
$$= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots\right) \pm j\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots\right)$$

It is also shown that

$$\cos x = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots\right)$$

and

$$\sin x = \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots\right)$$

It follows that

$$e^{+jx} = \cos x + j \sin x$$
$$e^{-jx} = \cos x - j \sin x$$
$$\cos x = \frac{e^{jx} + e^{-jx}}{2}$$
$$\sin x = \frac{e^{jx} - e^{-jx}}{2j}$$

The term $e^{\pm jx}$ represents a vector of unit magnitude, displaced from its zero or reference position by an angle of $\pm x$ radians.

Now, if $x = \omega t$, where $\omega = 2\pi f$ is the angular velocity of the rotating phasor in radians per second, and if f is the cyclic frequency in cps, the exponential $e^{\pm jx} = e^{\pm j\omega t}$ designates the locus of a circular path of unit amplitude, rotating with angular velocity of ω radians per sec, the positive direction of ω being that for which the phasor rotates in a counter-

clockwise direction. If the exponent is positive, the path is counterclockwise; if it is negative, the path is negative or clockwise.

The phase angle ϕ is taken into account by writing $x = \omega t + \phi$ and the amplitude is accounted for by an amplitude-multiplying factor E. Then,

$$Ee^{j(\omega t + \phi)} = Ee^{j\omega t}e^{j\phi} = E \cos (\omega t + \phi) + jE \sin (\omega t + \phi)$$

represents a phasor of magnitude E rotating in positive or counterclockwise direction, with angular velocity ω radians per sec, and having an initial phase displacement of ϕ radians, as shown in Fig. 5. In the same way, a phasor of magnitude E rotating in the negative or

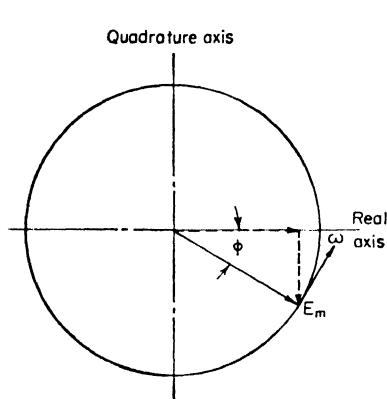



FIG. 5. Generating simple harmonic function by projecting on a reference axis the end of a phasor rotating with constant velocity about a fixed point. The dashed arrows show the rectangular components of the phasor, of magnitude E_m, represented by solid arrow.

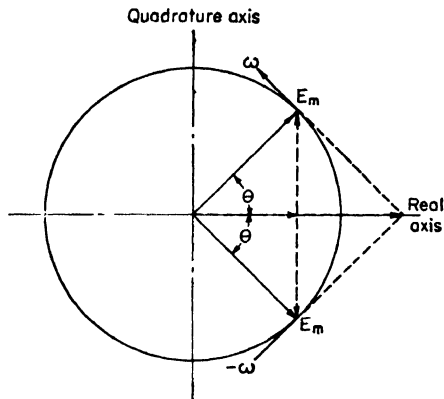



FIG. 6. Harmonic functions may also be generated by the resultant of two equal phasors rotating about a common point with velocities that are equal in magnitude but opposite in direction.

clockwise direction with constant angular velocity of $-\omega$ radians per sec, and having an initial phase displacement of θ radians, is expressed by the relation

$$Ee^{j(-\omega t + \phi)} = Ee^{-j\omega t}e^{j\phi} = E \cos (\omega t + \phi) - jE \sin (\omega t + \phi)$$

In the terminology of exponential functions, the real component represents the cosine function, as given by

$$E \cos (\omega t + \phi) = \text{Re}[Ee^{j(\omega t + \phi)}]$$

and the quadrature or imaginary component represents the sine function as given by

$$E \sin (\omega t + \phi) = \text{Qu}[Ee^{j(\omega t + \phi)}] = \text{Im}[Ee^{j(\omega t + \phi)}]$$

In engineering use, the real component is employed more frequently than the quadrature or imaginary component and the Re notation is understood; when this is the case, the Re notation is usually dropped as superfluous.

In addition to being expressible in terms of a single rotating phasor, harmonic functions may also be represented by a pair of phasors of equal magnitude E having the same initial phase displacement ϕ and having the same absolute value of angular frequency ω but with different sign, and hence rotating in opposite directions. As shown in Fig. 6, the cosine function may be expressed as one-half the vector sum of the oppositely rotating phasors, according to the relation

$$E \cos (\omega t + \phi) = \tfrac{1}{2}(Ee^{j\omega t}e^{j\phi} + Ee^{-j\omega t}e^{j\phi}) = Ee^{j\phi} \left(\frac{e^{j\omega t} + e^{-j\omega t}}{2}\right)$$

By forming the projection of the two vectors on the quadrature axis, the sine function, along the j axis, is expressed as

$$jE \sin (\omega t + \phi) = \tfrac{1}{2}[(Ee^{j\omega t}e^{j\phi}) + (-Ee^{-j\omega t}e^{j\phi})]$$

from which

$$E \sin (\omega t + \phi) = Ee^{j\phi} \left[\frac{e^{j\omega t} - e^{-j\omega t}}{2j} \right]$$

E above is an amplitude function and $e^{j\phi}$ represents a phase displacement. The exponential form of representing harmonic functions may be simplified by combining E and $e^{j\phi}$ to yield a pair of conjugate complex numbers to designate both the amplitude and the phase. This may be done by making use of

$$E' = |E|e^{j\phi} = E_+ + E_- = (A + jB) + (A - jB)$$

With this notation, $E_+ = (A + jB)$ and $E_- = (A - jB)$ are both complex quantities and are the conjugate of each other. The pair of conjugate complex quantities together make up the complex magnitude E, which expresses both the absolute value of the amplitude $|E|$ and the phase displacement ϕ. With such a notation, a rotating phasor may be represented by the simple expression

$$E = E'e^{j\omega t}$$

In this form, the phase angles are incorporated in the expression for E', which is the sum of the two conjugate complex numbers, but the phase angles do not appear explicitly. When required, however, the phase angles may be derived from the complex form of E' since

$$\phi = \tan^{-1} (B/A)$$

The exponential form of representing harmonic functions is particularly useful when the functions are to be differentiated or integrated. The rectangular or polar forms are each easily derived from the exponential form, which contains both sine and cosine functions, and hence the exponential form, when modified, lends itself also quite well to operations of addition and subtraction as well as to multiplication and division.

In the exponential form, the amplitude, frequency, and phase components may be readily separated. At the same time, by making use of a pair of conjugate coefficients, it is possible to combine the amplitude and phase into a pair of conjugate terms which, in the equations, may be handled as a single term. Thus the E coefficients are complex, but they appear in pairs and their sum is a real number. The exponential representation of a rotating phasor includes both the real (or cosine) term and the quadrature (or sine) term, either one of which may be selected merely by choosing the projection on the real or the quadrature axis, respectively.

39. Nature of Harmonic Analysis. By a technique first disclosed by Fourier, any periodic wave form can be analyzed into, or can be synthesized from, a series (which may contain an infinite number of terms) of sinusoids or harmonic functions. The frequencies of the terms in the series are integral multiples of some fundamental frequency, and the amplitudes and phases of the harmonic components—their spectra in other words—are uniquely related to any given wave form. The practical application of the Fourier technique depends, of course, upon being able to determine the amplitude and phase spectra for each and every term in the series representing the given wave form. This technique is accomplished by that branch of mathematics known as *harmonic analysis*. Reciprocally related Fourier transformations in harmonic analysis make it possible to specify a given function either in terms of its amplitude-time function in the time domain, or by means of its amplitude and phase spectra in the frequency domain.

By extension of the Fourier technique, any nonrecurrent phenomena may be considered as the limiting case of periodic phenomena with infinite period. By a limiting process, the methods of harmonic analysis for periodic phenomena may be extended to apply to nonrecurrent or singularity functions. Again, the practical application of the Fourier integral technique depends upon being able to determine the amplitude and phase spectra for the transient phenomena of specified wave form. This is accomplished by generalizing the technique of harmonic analysis, and again suitable transformations may be made from the time domain to the frequency domain and vice versa.

Still further generalization of harmonic analysis makes it possible to characterize a random probability-distribution function by means of a *correlation function* which is equivalent to the power spectrum of the random function.

At the present time a great deal of activity is going on in the development of information theory, in which the theory of probability plays as significant a role as does generalized harmonic analysis, especially as this is applied to stationary time series. We shall limit the treatment here to a brief discussion of Fourier series and integrals, leaving the user of this compendium to refer to specialized texts for treatment of more recently developed and more advanced topics in this field.

40. Fourier-series Analysis. Any recurrent, periodic, single-valued wave form whose nonrecurrent interval is T may be expressed as the sum of a series of sinusoids whose frequencies are integral multiples of the lowest or fundamental frequency $f = T/2\pi$. A constant term, which may be regarded as a harmonic function of zero frequency, may be required in such a series.

As an example of Fourier's method, Fig. 7 shows a rectangular, periodic wave form (top) with the required amplitude spectrum (center) and phase spectrum (bottom).

The general Fourier-series expansion for any periodic function may be written in the alternative forms:

$$e(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos n\omega t + \sum_{n=1}^{\infty} B_n \sin n\omega t$$

$$= \sum_{n=-\infty}^{\infty} C_n e^{jn\omega t}$$

where $n = 0, 1, 2, 3, \ldots, n$ is the order of the harmonic. The amplitudes of successive terms in either and both forms of the series must decrease with increasing harmonic order for the series to converge to a finite limit; this condition is fulfilled for wave forms ordinarily encountered in radio communication.

Amplitude-time plot of wave of rectangular form

Amplitude-frequency plot of rectangular wave form, with frequency scale expressed in terms of multiples of the fundamental frequency

+π
Phase 0
−π

Phase-frequency plot of rectangular wave form, with frequency scale identical to that for the amplitude-frequency plot

FIG. 7. Square wave plotted against time with its corresponding amplitude-frequency and phase-frequency plots.

For either of the above expressions to specify a periodic function properly, it is necessary to know the numerical values of the amplitudes of the harmonic functions of the Fourier series, *i.e.*, we need to know the values of A_0, A_n, and B_n, or of C_n for all values of n. To evaluate these coefficients, use is made of the following trigonometric integrals:[1]

1. The average value of a single sine or cosine term, when taken over a complete period, is zero.
2. The average value of the product of two sines or cosines or of a sine and a cosine of commensurable but numerically unequal frequencies, taken over any complete period of the product, is zero.
3. The average value of the product of two sines or cosines of the same frequency, taken over any complete period, is one-half the cosine of the difference between the phases of the two factors. For the product of a sine and cosine, the average is one-half the sine of the angle by which the sine term leads the cosine term.

If $e(t)$ is an arbitrary periodic function of period T, so that $e(t) = e(t \pm T)$, it is expressible as a Fourier series of harmonic functions. If the function $e(t)$ is specified, either analytically or graphically, the Fourier coefficient.

[1] Franklin, P., "Differential Equations for Electrical Engineers," Wiley, 1933.

may be determined from the relations for the line spectrum

$$A_0 = \frac{1}{T}\int_t^{t+T} e(t)\,dt$$

$$A_n = \frac{2}{T}\int_t^{t+T} e(t)\cos n\omega t\,dt$$

$$B_n = \frac{2}{T}\int_t^{t+T} e(t)\sin n\omega t\,dt$$

$$C_n = \frac{1}{T}\int_t^{t+T} e(t)\epsilon^{-jn\omega t}\,dt$$

where $T = 2\pi/\omega$ is the fundamental period, or the period of the lowest frequency or fundamental term. In the above expression, C_n is a complex quantity, related to A_n and B_n by the equations

$$C_0 = A_0 \qquad C_{+n} = \tfrac{1}{2}(A_n + jB_n) \qquad C_{-n} = \tfrac{1}{2}(A_n - jB_n)$$

When the Fourier series represented by $e(t)$ is inserted into the integral expressions for the Fourier line-spectrum coefficients, all terms, when integrated over a complete period, become zero except those for the nth term. Hence, for each value of n it is possible to determine the Fourier coefficients.

Any pair of conjugate terms in C_n specifies an n-term frequency component of the Fourier series, since

$$C_{+n}e^{jnx} + C_{-n}e^{-jnx} = A_n \cos nx + B_n \sin nx$$

The absolute value or magnitude of the nth term is

$$C_n = (A_n^2 + B_n^2)^{1/2}$$

and the phase angle for the nth term is

$$\phi_n = \tan^{-1} B_n/A_n$$

By means of the C_n and ϕ_n terms, the amplitude and phase may be plotted for the nth harmonic to produce a line spectrum which is characteristic of that Fourier series and the wave form it represents. In the exponential form of the expression for harmonic functions, the C_n terms occur in conjugate pairs; from such a conjugate pair, the amplitude and phase of the nth term may be readily derived, but the phase properties do not appear explicitly.

For periodic functions $e(t)$ of period T, the function is analyzed into its line or discrete amplitude- and phase-frequency spectra by means of the relation

$$C_n = \frac{1}{T}\int_0^T e(t)\epsilon^{-jn\omega_1 t}\,dt$$

where ω_1 is the angular frequency (in radians per second) corresponding to the period T (in seconds). The function $e(t)$ is synthesized from its amplitude- and phase-frequency spectra by the transformation

$$e(t) = \sum_{n=-\infty}^{\infty} C_n e^{jn\omega t}$$

41. Average and Effective Values of Periodic Functions. If $e(t)$ is a periodic function of time with period T so that $e(t) = e(t + T)$, the average value of the function over the time interval of a complete period T is

$$E_{av} = \frac{1}{T}\int_t^{t+T} e(t)\,dt$$

The average value of a cyclic function which is symmetrical with respect to the zero axis throughout the cycle is zero. Over any integral number of complete cycles, the average value of a sine or cosine function, therefore, is zero.

In speaking of the average value of a sine or cosine function, the electrical engineer refers to the average value of either the positive or the negative half of the cycle, the

absolute value of which is

$$E_{av} = \frac{1}{\pi}\int_0^{\pi} E_m \sin \omega t\, d(\omega t) = \frac{2E_m}{\pi} \cong 0.6366198 \ldots E_m$$

where E_m is the maximum value of the harmonic function.

In many cases, certain electrical effects are proportional to the square root of the average value of the square of the function throughout the cycle. This value is known as the *effective value*, or the *root-mean-square* (rms) value from the fact that it is found by first squaring the function, determining the mean value of this squared value, and extracting the square root of the mean value to determine the end result.

Thus, in general, the effective value of a function of time $e(t)$ is given by the relation

$$E_{rms} = \left\{\frac{1}{T}\int_t^{t+T} [e(t)]^2\, dt\right\}^{1/2}$$

If the time function is a harmonic function such that $e(t) = E_m \sin \omega t$, the rms value of the function over a complete cycle or period is

$$E_{rms} = \left[\frac{1}{2\pi}\int_0^{2\pi} (E_m \sin \omega t)^2\, d(\omega t)\right]^{1/2} = \frac{E_m}{\sqrt{2}} \cong 0.7071068 \ldots E_m$$

By applying the equation for the effective or rms value to a periodic wave form expressible by the Fourier series,

$$e(t) = E_0 + E_1 \sin (\omega t + \phi_1) + E_2 \sin (2\omega t + \phi_2) + \cdots + E_n \sin (n\omega t + \phi_n) + \cdots$$

the effective value of a periodic recurrent function is given by

$$e(t)_{rms} = [E_0^2 + \tfrac{1}{2}(E_1^2 + E_2^2 + \cdots + E_n^2 + \cdots)]^{1/2}$$

42. Aperiodic Functions. The Fourier series represents a recurrent time function whose energy is represented by a number of discrete and uniformly spaced frequency components, the sum of whose amplitude and phase components constitutes a pair of line spectra. In general, the frequency components of a periodic wave may be expected to be infinite in number, and each is represented by a trigonometric function of time, mathematically defined as existing throughout all intervals of time from $t = -\infty$ to $t = \infty$. Since the wave form thus described by a Fourier series is expressed in terms of functions which are defined as having existed for all values of time, the Fourier series—or its equivalent representation as a pair of line spectra—can be used to represent only steady-state phenomena.

When we come to deal with aperiodic or singularity functions, we find that such functions can be analyzed or synthesized by a process similar to that employed for representing periodic functions, although there are some differences. For aperiodic or transient phenomena, the energy represented by a pair of amplitude and phase spectra must be continuous, whereas they were line spectra in the case of periodic phenomena. There are no recurrent features in an aperiodic function, which, accordingly, may be regarded as the equivalent of a periodic function with infinitely long cycle of recurrence. As the period becomes infinitely long, the frequency separation between the discrete components in the line spectra approaches zero. Accordingly, the aperiodic function will be a continuous function of angular frequency ω rather than a function of discrete frequency intervals $n\omega_1$, where ω_1 is the angular frequency of the lowest-frequency component.

43. Fourier Integral. From the discussion on aperiodic functions, it would seem logical that aperiodic phenomena might be derivable from the Fourier-series expression of periodic phenomena by means of a limiting process in which the series form is, in some way, converted to the integral form.

To see how such a limiting process may be effected first write the expression for the periodic function as

$$e(t) = \sum_{n=-\infty}^{\infty} C_n \epsilon^{jn\omega t} = \sum_{n=-\infty}^{\infty} \left(\frac{C_n}{\omega_1}\right) \epsilon^{jn\omega t}\omega_1$$

where, now, ω_1 is used to designate the frequency interval between the successive discrete frequency components. If T is the period, f_1 the fundamental frequency in cps, and ω_1 is fundamental angular frequency in radians per second, then $T = 2\pi/\omega_1$, and a complete cycle of recurrent phenomena occurs between the limits of $T = -\pi/\omega_1$ and $T = +\pi/\omega_1$. Hence, the expression for the amplitude and phase spectra may be written as

$$C_n = \frac{1}{T}\int_t^{t+T} e(t)\epsilon^{-jn\omega_1 t}\,dt = \frac{\omega}{2\pi}\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} e(t)\epsilon^{-jn\omega_1 t}\,dt$$

In the limiting process, as $T \to \infty$, it follows that $f_1 \to 0$ and $\omega_1 \to 0$. Moreover, since $C_n \propto 1/T$, the amplitudes of any given harmonic components approach zero as the period approaches infinity, although the ratio C_n/ω_1 may be expected to approach a finite limiting function as the period increases to infinity. No matter what the line spacing ω_1 may be, since n is an integer it follows that the spacing between successive spectral lines, or the frequency increment between successive lines, is equal to the lowest or fundamental frequency, and we may write $\Delta(n\omega_1) = \omega_1$. With this notation $e(t)$ takes the form

$$e(t) = \sum_{n\omega_1=-\infty}^{\infty} \left(\frac{C_n}{\omega_1}\right) \epsilon^{jn\omega_1 t}\,\Delta(n\omega_1)$$

and the function representing the amplitude and phase as a function of frequency becomes

$$\left(\frac{C_n}{\omega_1}\right) = \frac{\omega_1}{2\pi}\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} e(t)\epsilon^{-jn\omega_1 t}\,dt$$

Now, to effect the limiting process, let $T \to \infty$, $\omega_1 \to 0$, $n \to \infty$, and $(n\omega_1) \to \omega$. Also, in the limit, we define $\Delta(n\omega_1) = d\omega$ and also $(C_n/\omega_1) = C(\omega)$. When such a limiting process is effected, the aperiodic function is expressible by

$$e(t) = \int_{-\infty}^{\infty} C(\omega)\epsilon^{j\omega t}\,d\omega$$

where

$$C(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e(t)\epsilon^{-j\omega t}\,dt$$

By suitable transformations, $e(t)$ may be transformed into $C(\omega)$ and vice versa. Thus,

$$C(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \epsilon^{-j\omega t}\,dt \int_{-\infty}^{\infty} C(\omega)\epsilon^{j\omega t}\,d\omega$$

and

$$e(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \epsilon^{j\omega t}\,d\omega \int_{-\infty}^{\infty} e(t)\epsilon^{-j\omega t}\,dt$$

The above expression for $e(t)$ is called the *Fourier integral* representation of the function $e(t)$, or the *inverse transform* of $C(\omega)$. The expression for $C(\omega)$ is the *Fourier transform* of $e(t)$ because it transforms a time function $e(t)$ into its equivalent frequency function $C(\omega)$. The expression for $e(t)$ synthesizes a given time function from its amplitude and phase spectra, which are now continuous functions of frequency; it represents the function in the time domain. The expression for $C(\omega)$ represents the

function in the frequency domain and analyzes the given function in its amplitude and phase spectra. The two equations may also be regarded as representing a pair of integral equations and their solutions.

44. Random Functions. Random functions are devoid of any fixed or precisely predictable relationships between the dependent and the independent variables. Therefore, it is not possible to write an expression for the amplitude of a random function. It is possible, however, to state the probability with which the amplitude will lie between two specified values of the dependent variable for some value of the independent variable. Random functions are treated analytically by means of statistics, which is based on the theory of probability.

45. Noise. In an acoustic sense, noise may be defined as any extraneous sound tending to interfere with the proper and easy perception of those sounds or signals which it is desired to receive. In this sense any undesired signal may be regarded as noise when mixed with the desired signal. Many disturbances of an electrical nature produce noise in radio receivers, so that the term has a much broader connotation in communication engineering than that given by the above definition. More generally, then, noise may be defined as some random process which modifies the signal in a predictable manner. The principal effect of noise is to interfere with the desired signal and make it less easy to perceive; it limits range and sensitivity.

Noise in radio systems may be of two general kinds. *Random noise* is noise due to a large number of individual, but not necessarily identical, overlapping disturbances, whose occurrence has no specifiable regularity. Nonrandom noise is noise that exhibits more or less definite regularities and hence assumes more nearly the character of interference from some other signal.

An important property of random noise is that the addition of two or more random-noise signals produces a resultant random-noise signal whose power is equal to the sum of the power of each of the component noise signals. Power, in this sense, is defined as the integral of the square of the amplitude.

Random noise includes that due to: (1) shot noise, resulting from fluctuations in electron emission in circuits involving thermionic emission; (2) incidental noise of atomic origin in electron tubes, including that due to ionization, production of secondary electrons, emission of positive ions, and fluctuations in these phenomena; (3) flicker noise, resulting from variations of the total space current in a thermionic emitter, as a result of changes in evaporation, diffusion, chemical action, structural rearrangement, and gas-ion bombardment throughout areas of the cathode surface which are large compared with that involved in the emission of a single electron; (4) "static" electrical charges, atmospherics, and interstellar noise and interference; (5) contact or breakdown noise, resulting from the breakdown of insulation, or the making and breaking of minute electrical conducting paths in components of the system; (6) dirt and grain-size noise, which are attributable to minute irregularities in the structure of equipment or components, and which may be innate in the materials employed, or developed in processing or handling.

Nonrandom noise is not of random nature and has characteristic wave shapes of its own. Nonrandom noises that may be encountered in electrical-communication systems include: (1) noise from rotating or vibrating electrical equipment, *e.g.*, motor commutators, electric razors, and vacuum cleaners; (2) ignition noise from ignition systems; (3) hum from electric power lines, of fundamental and harmonic frequencies; (4) extraneous signals, such as harmonics of stations operating at a submultiple frequency of that to which the receiver is tuned, or signals from stations operating at "sum and difference" frequencies of local oscillators within radio receivers; (5) noise due to some irregularity that is periodic at the frequency of rotation of mechanical equipment, commonly called "wow"; (6) mechanical vibrations of components in the radio receiver, commonly called "microphonics"; (7) feedback and oscillation within the system, called "howl" when occurring at an audio frequency, or "motorboating" when occurring at a subaudible frequency; (8) spurious responses of the equipment, and other noises.

46. Information Theory. Since the beginning of the Second World War, considerable progress has been made in the development of a theory of information which takes full cognizance of the unpredictability of messages, and applies probability

theory to the specification of messages. Progress in this field has been so rapid and so extensive that only the most elementary highlights can be set forth here.[1]

Communication is carried on by the successive selection and arrangement of prearranged or generally understood physical symbols. In a manner that depends upon the properties of the communication system, such signals are produced by the sender and are brought to the attention of the receiver. The proper selection and sequential arrangement of such symbols or signals are capable of conveying intelligence in accordance with the meanings attached to the symbols. Since a sequence of symbols is required to convey information, it is evident that one of the characteristics of intelligence is that it may be expressed as a single-valued time function. From a practical point of view, this is a most fortunate result, for it is this characteristic property of intelligence that permits information to be transmitted economically by electrical means.

In the process of communication, it is clear that two entirely distinct operations are involved. One of these deals with the proper selection and sequential arrangement of the physical symbols; the other deals with the meanings attached to the symbols. Information theory is concerned with the first of these two phases and with such matters as definitions and measurement of information, the philosophy of information, random processes, probability theory, correlation, prediction, filtering, storage, modulation, and like topics, all of which are equally valid without regard to the meanings attached to the symbols of communication. It is this phase of the broad topic with which communications engineers are primarily—if not exclusively—concerned. That phase of communication dealing with the meanings attached to the symbols is, in fact, semasiology, which is the province of semanticists and linguists.

Information is conveyed by the successive selection and arrangement of a number or sequence of symbols; it is measured by the number of symbols needed to convey the desired idea or intelligence. Fundamental to the concept of information is the idea that, prior to its reception, any message is uncertain; the more uncertain we are about the composition of the message, the greater will be the amount of information needed to remove the uncertainty regarding the unpredictability as to what has been sent. It is, therefore, the unpredictable nature of the message that makes it convey information. According to this point of view, whatever their source, messages are regarded as random time-varying functions. It is usually assumed that the statistical character of the random time functions remains invariant with time.

To be useful in electrical communication, any measure of information must be correlated with the notion that the amount of information increases as the number of selections or choices from a set of possibilities is increased. The simplest possible choice is the selection between two equally likely choices. It appears reasonable, therefore, to define the unit of information—the amount, not the meaning—in terms of such equally likely selections between M possibilities. We also recognize that n independent selections (*i.e.*, selections which do not affect one another) of this type constitute n units of information. Hence the basic unit of information is defined as the amount of information associated with a single selection between M equally likely choices; invariably the choice is made between only two possibilities at one time, and for such a binary choice, of course, $M = 2$. The name "bit" from "binary digit" has been almost universally adopted to designate the unit of information, although the term "binit," having the same derivation, has also been suggested.

When a selection is made from M choices, not necessarily equally likely, we shall say that an M-order selection has been made. It will be convenient at times to retain this generality, although nearly all work in information theory has dealt with binary selections for which $M = 2$. When a series of n selections is made in M-order

[1] For fuller details, the reader is referred to "A Bibliography of Information Theory (Communication Theory—Cybernetics)," by F. L. Stumpers (published Feb. 2, 1953, by the Research Laboratory of Electronics, M.I.T., Cambridge, Mass.), listing approximately 800 contributions under a dozen different headings. For subsequent work, the reader is referred to the "Symposia of Information Theory" issued annually by the Institute of Radio Engineers.

selections, we say that an n-order choice has been made. When the choices are, a priori, equally likely, the selection is called an equally likely selection.

To illustrate the relationship between information and a sequence of selections, consider the diagram of Fig. 8 in which, proceeding from left to right, an M-order selection is made at each choice, and there are n choices. In the diagram shown, we deal with binary selection, so $M = 2$, and there are four sequences, so that $n = 4$. Beginning at the initial point, in the first step we make a selection between two alternatives (a 0 or a 1) which lead to either the top or the bottom half of the remaining portion of the diagram. In the second choice, we select an upper (0) or lower (1)

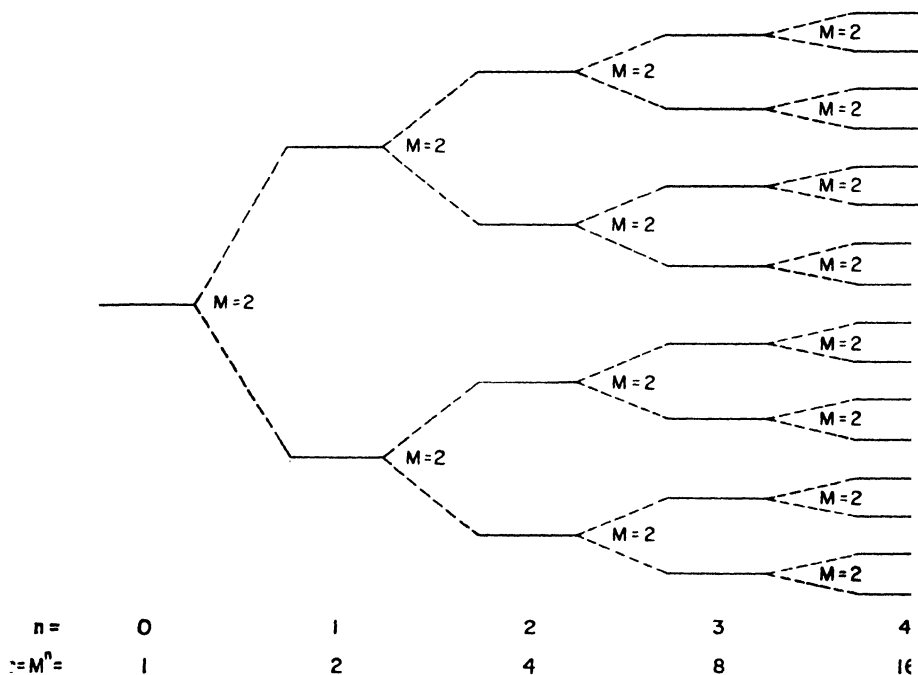



FIG. 8. Relationship between information and a sequence of selections. At each step the number of possible selections or choices is 2 ($M = 2$); the number of sequence steps is designated as n. The total number of possible sequences is $c = M^n$.

group in the half previously chosen. In the third choice we select an upper (0) or lower (1) subgroup in the group previously chosen, and so on until a final selection has been made from n possibilities. A new unit of information is added each time a choice is made, for the selection excludes all irrelevant possibilities and more accurately and definitely specifies a given selection, thereby removing some uncertainty regarding the message. As the number of choices increases linearly, the number of possibilities from which a selection is possible increases geometrically, or in an exponential sense. Hence, if c is the total number of possible sequences in an M-order selection with n choices all being equally probable, then

$$c = M^n$$

If information increases linearly with the number of selections, which is in accordance with our intuitive feeling, the quantity of information H must be dependent upon c in such a way that

$$H = f(M^n) = f(c)$$

The amount of information H is linearly related to the number of selections n if the function f is such that

$$H = n \log M = \log c$$

In this case of binary selection ($M = 2$) we have been able to select any one of

$$c = 16 \text{ possibilities}$$

all equally likely, with $n = 4$ choices. Since four selections were required, we have 4 bits of information represented by the process described and shown in Fig. 8.

The base of the system of logarithms has not been stated in the result given above. This is not of fundamental importance, because logarithmic measurements in either of two systems of measurement (or to either of two bases) differ merely by a constant, which can easily be taken into account by a proportionality factor. The selection of a base is, then, largely a matter of convenience. The simplest results obtain when logarithms to the base M are used, for in this case $\log_M M = 1$, and for the case under discussion, then $H = n$.

The result given above is strictly true for binary selection if M is a power of 2; in such a case H then becomes an integer. If M is some number other than a power of 2, then, for binary selections, the number of elementary selections required to specify the desired choice will not be an integer; it will be a number equal to the logarithm of either the next higher or the next lower power of 2. The corresponding amount of information must, accordingly, lie between these two limits, and we must then determine the significance of amounts of information H which are not expressible by integers. The equation given above is correct, even for nonintegral values of H (that is, for M not a power of 2) provided we regard and interpret H to be the average value of the information conveyed by a large number of selections, all of which are equally likely, rather than the total information as given above.

In dealing with long messages in which many choices are made, it is often convenient to have a quantity that represents the average amount of information per M-order selection rather than, as has been done above, to determine the total amount of information corresponding to the whole message or total number of sequences of selections. The average information H_{av} may be defined as the total amount of information given by a very long message H divided by the number of choices n in an M-order selection. When all the selections in the message are equally likely and independent and when, in addition, M is a power of 2, the quantity H_{av} is just equal to the information actually given by each selection, that is,

$$H_{av} = \frac{1}{n} \log_2 c = \log_2 M$$

If M is not a power of 2, this equation represents the average amount of information taken over a very long sequence of selections.

It is now necessary to take into account the probability that certain messages or selections are made, for messages having a high probability of being chosen will require fewer selections than messages having a low probability of being chosen. Furthermore, we feel instinctively that a message of low probability—and hence more selections—contains more information than a message of high probability requiring few selections.

To obtain a mathematical expression for this state of affairs, let $P(i)$ be the probability of the ith message. The number of binary selections required to convey such a message will be some integer $B_S(i)$ whose magnitude will be close to $-\log_2 P(i)$. But $P(i)$ is the probability of making a selection in the last subgroup through a process of successive halving the entire ensemble of messages $B_S(i)$ times, as already described. In other words, for a relatively short or small number of selections, $P(i) \cong 2^{B_S(i)}$, and by making the number n of M-order selections sufficiently large, the percentage difference between $-\log_2 P(i)$ and $B_S(i)$ can be made as small as we choose. In the limiting case of an infinite number of selections (*i.e.*, $n \to \infty$),

$$B_S(i) = -\log_2 P(i)$$

Consider now a sequence of Q selections of messages, each message of which is assumed to consist of n M-order selections so that the total sequence represents nQ selections. If the number Q is made sufficiently large, the ith message in the sequence will have a frequency as close to $P(i)$ as we may choose. The average number of binary selections required to select one message will be the mathematical expectation of B_S, or $E(B_S)$. If c is the number of different sequences that can be formed, then

$$E(B_S) = -\sum_{i=0}^{c-1} P(i)B_S(i)$$

The average amount of information produced by each M-order selection is then the ratio of the expectation of B_S to the number of selections n as $n \to \infty$. That is,

$$H_{av} = \lim_{n\to\infty} \frac{E(B_S)}{n} = \lim_{n\to\infty} -\left(\frac{1}{n}\right)\sum_{i=0}^{c-1} P(i)\log_2 P(i)$$

This limiting process yields[1] a result for the average amount of information per M-order selection of

$$H_{av} = -\sum_{k=0}^{M-1} p(k)\log_2 p(k)$$

where $p(k)$ is the probability of the kth choice of M, and n_k is the number of times the kth choice is selected in the ith message of n selections. A negative sign is used here because, although the average amount of information per M-order selection is positive, the individual probabilities are less than unity, so that their logarithms are negative numbers; these are transformed into positive numbers by the minus sign.

47. Band Width and Noise. A practical problem in communication theory has to do with the rate at which a function to be transmitted should be sampled so that it may retain its essential characteristic information. Or we may wish to know how many amplitudes must be distinguished from one another so that the communication system serves its intended purpose adequately. An infinite number of amplitudes may be theoretically desirable, but in general this requires an infinite band width and can be achieved only in the absence of noise. Such ideal conditions are not encountered in practice. The number of distinguishable amplitudes is limited by the amount of noise, or, more correctly, by the ratio of signal-to-noise power. Furthermore, practical or economic or legal factors may dictate an upper band width so that we are forced to transmit and receive a function that differs somewhat from the original signal.

The band width of the transmission system is related to the number of samplings we can make of the original signal per unit time. If we are limited to a transmission band width of f_B cps, this band width will be fully needed when the original signal is sampled at instants spaced $1/2f_B$ time units (seconds) apart.

That is,[2] the specification of an arbitrary wave of duration T and maximum frequency component f_B requires $2f_BT$ measurements; the Fourier series yields $2f_B$ coefficients and an upper frequency limit of f_B cps.

If the average random noise power is N, and if the average signal power is S, then $(S + N)$ will be the average total power of signal and noise when both are mixed. The effective values of the amplitudes are the square roots of the respective power values. Therefore, the ratio of the total power to noise power is $(S + N)/N$, and the ratio of the amplitudes of signal and noise to noise is

$$\frac{\sqrt{S+N}}{\sqrt{N}} = \sqrt{\frac{S+N}{N}} = \left(1 + \frac{S}{N}\right)^{1/2}$$

[1] Fano, R. M., "The Transmission of Information," *Technical Report*, 65, M.I.T. Research Laboratory of Electronics.

[2] Bennett, W. R., Time Division Multiplex Systems, *Bell System Tech. J.*, **20** (2), 199, April, 1941.

If it is assumed (as is true to a close approximation) that a change in signal amplitude cannot be instantly recognized when it is less than the noise amplitude, whereas a signal may be recognized if its amplitude is equal to or greater than that of the noise, then the number of distinguishable signal amplitudes in the presence of noise is

$$\delta = k\left(1 + \frac{S}{N}\right)^{1/2}$$

where the factor k depends at least in part on system reliability, but is often in the vicinity of unity.

If the function containing signal and noise is sampled every $1/2f_B$ sec, then in time T there will be $2f_B$ samplings, as well as a like number of amplitudes of the function which are independent of each other. It follows that the number of distinct signals in T sec from which one makes a choice under the stated conditions is

$$\delta^{2f_BT} = k\sqrt{\left(1 + \frac{S}{N}\right)^{2f_BT}} = k\left(1 + \frac{S}{N}\right)^{f_BT}$$

The amount of information—the number of bits—sent in time T, when the band width of the communication facility is f_B in the presence of a signal-to-noise power ratio of S/N, will be

$$H = \log_2 \delta^{2f_BT} = kTf_B \log_2\left(1 + \frac{S}{N}\right)$$

Compare this result with Hartley's law, which states that the amount of information transmitted over a channel of band width f_B cps when used for T sec is proportional to the product of the band width and the transmission time, or $H = kTf_B$.

Before proceeding, let us interpret the significance of the expression above. This equation shows that, in the presence of a given noise power N, the total quantity of information that may be transmitted depends upon a number of factors. If these are independently variable, we have a certain amount of freedom in designing the communication system to convey a given amount of information in the presence of a stated amount of noise. In other words, in the presence of a given noise power, the total quantity of information may be kept constant, even though the individual factors comprising the information may be varied. That is, to transmit a specified quantity of information in the presence of a given amount of noise, the band width, transmission time, or signal power may be changed, within limits, to yield the most satisfactory communication circuit. To increase the amount of information transmitted, it is most effective to increase f_B or T, since H is directly proportional to band width or transmission time. However, if there are restrictions on f_B and T, we may still achieve some gain by increasing the power of the signal. The gain in information transmitted by multiplying the signal power by a certain factor is less than that obtained by multiplying either the band width or the time of transmission by the same factor.

The above equation shows that we can increase the amount of information by: (1) increasing the band width, (2) increasing the time of transmission, (3) increasing the signal power, (4) decreasing the noise power, or (5) increasing the signal-to-noise power ratio.

The relations between band width, transmission time, and signal power of signals transmitted in the presence of noise have been investigated by Tuller,[1] who shows the fallacies of earlier concepts of information transmission.

The type of communication system considered by Tuller is one in which the frequency band width of the transmission link f_c is wider than the frequency band width of the over-all communication system f_B or at least is different from it, in general. Moreover, N is the rms amplitude of the noise when this becomes mixed with the signal, and C is the maximum signal amplitude, measured before nonlinear operations in the receiver, such as detection.

[1] Tuller, W. G., Theoretical Limitations on the Rate of Transmission of Information, *Proc. IRE*, **37** (5), 468, May, 1949. Also doctoral thesis, Department of Electrical Engineering, M.I.T., June, 1948.

With these symbols, the quantity of information to the input of the receiver is

$$H_{\text{in}} = 2kTf_c \log\left(1 + \frac{C}{N}\right)$$

whereas the amount of information at the output of the system is

$$H_{\text{out}} = 2kTf_B \log\left(1 + \frac{S}{N}\right)$$

where k is a proportionality constant. The fact that the receiver cannot add any information but may lose some information establishes the relationship $H_{\text{out}} \leqq H_{\text{in}}$, which, in terms of the symbols used above, yields the general result

$$\left(1 + \frac{S}{N}\right) \leqq \left(1 + \frac{C}{N}\right)^{f_c/f_B}$$

In a circuit in which there is little noise, $S/N \gg 1$ and $C/N \gg 1$ and the general result may be simplified to

$$(S/N) = (C/N)^{f_c/f_B}$$

Now, $(1 + S/N) = s_{\text{out}}$ must be the number of possible allowed states of the receiver output at any one time, and $(1 + C/N) = s_{\text{in}}$ is the number of possible allowed states at the receiver input for any one instant, so that the general result may be expressed as

$$s_{\text{out}} \leqq (s_{\text{in}})^{f_c/f_B}$$

For coded transmission equality in the above expression can be achieved, in which case

$$\left(1 + \frac{S}{N}\right) = \left(1 + \frac{C}{N}\right)^{f_c/f_B}$$

whereas for uncoded transmission

$$\left(1 + \frac{S}{N}\right) \leqq (f_c/f_B)\left(1 + \frac{C}{N}\right)$$

The results show that wide-band modulation systems using uncoded transformations are inherently inefficient in their spectrum utilization, and that coding may be used as desired to gain in one parameter at a sacrifice in some other or others, without loss in efficiency.

Finally, we may note that channel capacity in a communication system has been defined as the time rate at which information may be transmitted; thus, if κ is the channel capacity, then in terms of the symbols already used,

$$\kappa = \frac{H}{T} = kf_c \log\left(1 + \frac{S}{N}\right)$$

Thus channel capacity depends upon band width and upon the signal-to-noise ratio.

48. Ideal and Distortionless Transmission. It is desirable that the transmission and energy conversions in a communication system, from transmitter to receiver, be carried out so that, on the whole, the system is free from defects and distortions that prevent it from re-creating essentially the subjective sensations, impressions, or effects at the receiving end that were initially produced at the transmitter. Distortion may be produced by noise or by a change in wave form and may prevent this desirable objective from being realized. Distortion defects are most easily observed in terms of deviations from the original signal, and the original signal can be perfectly reproduced at the receiving end only by a system having ideal over-all transmission.

We may define perfect or ideal transmission as that for which the intelligence function re-created at the receiver is identical in all respects to the originally produced information function at the transmitter. The condition imposed on ideal transmission

systems may thus be expressed mathematically by the relationship

$$i_2(t) \equiv i_1(t)$$

The extent to which $i_2(t)$ at the receiving end differs from $i_1(t)$ at the transmitting end is a measure of the over-all deviation from the conditions of ideal transmission.

The requirement for ideal transmission, as given above, is more stringent than is usually necessary or economically feasible. Although perfect transmission is never precisely achieved, and does not need to be, nevertheless the concept of perfect transmission as defined above is exceedingly useful in establishing a criterion by which the performance of any communication system may be judged against the theoretically ideal goal.

For engineering purposes, the concept of distortionless transmission is a more useful one than perfect or ideal transmission. For distortionless transmission, the communication system should produce at the receiving end the same wave form (*i.e.*, the same intelligence-time function) as that originally produced at the transmitting end. Distortionless transmission allows for the finite time of passage of the signal through the transmission medium; usually it also allows for a change in scale or magnitude. Thus, for distortionless transmission, we may establish the criterion that

$$i_2(t - t_d) \equiv Ai_1(t)$$

where A is a scale or magnitude factor, and t_d is the time required for the signals to be transmitted.

Distortionless transmission differs from ideal transmission in that the sensation-perceived energy at the receiver may be greater than, equal to, or less than the corresponding energy at the transmitting end, suitable corrections being made, when necessary, by amplification or attenuation at the receiver. The received signal is also subject to a time delay or time displacement, since transmission does not occur instantaneously. Provided it is not excessive, a finite time of delay is usually no hindrance in most communication systems, but (as shown in Sec. 49) the time delay for the various frequency components of a complex signal must be properly related if the wave form of the signal is to remain unchanged. In some radio systems, of which radar is a notable example, the time delay for transmission is used to good advantage; in other systems—such as tape recording—the time delay may be very great before the signal is re-created.

49. Spectrum Response of Distortionless System. If we establish the criterion for distortionless transmission, as developed in Sec. 48, the spectrum response of a distortionless system can be easily derived. In making the following derivation, it is assumed that the system, as a whole, is linear, so that an output voltage or current of a signal is directly proportional to the input voltage or current that produced it. It is also assumed that the output of the system depends only upon the character of the input signal and the characteristics of the transmission properties of the system.

Assume an arbitrary input voltage whose value is expressed by the Fourier integral

$$e_1(t) = \int_{-\infty}^{\infty} C(\omega)e^{j\omega t}\,d\omega$$

where $C(\omega)$ specifies the amplitude and phase properties of the system. The term may be written as

$$C(\omega) = G(\omega)e^{j\phi(\omega)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e_1(t)e^{-j\omega t}\,dt$$

if $G(\omega)$ is the amplitude spectrum and $e^{j\phi(\omega)}$ is the phase spectrum of the input signal. This signal $e_1(t)$ is passed through the system, during which it will be delayed by a time interval τ and will be subject to modifications in its amplitude and phase spectra by an amount depending upon the over-all system transmission characteristics. Let $\tau(\omega) = A(\omega)e^{-j\theta(\omega)}$ be the over-all transmission properties of the system, in which $A(\omega)$ is the amplitude spectrum and $e^{-j\theta(\omega)} = \underline{/-\theta(\omega)}$ is the phase spectrum of the system.

Now, if $Y_{12}(\omega)$ is the transfer admittance of the over-all system—which may be resolved into its amplitude and phase spectra—then the output current $i_2(t)$ as a result of input voltage $e_1(t)$ is

$$i_2(t) = \int_{-\infty}^{\infty} C(\omega) Y_{12}(\omega) e^{j\omega(t-\tau)}\, d\omega$$

and the output voltage is

$$e_2(t) = \int_{-\infty}^{\infty} C(\omega) T(\omega) e^{j\omega(t-\tau)}\, d\omega$$

where $T(\omega) = Y_{12}(\omega) Z_R(\omega)$, if Z_R is the load impedance across which $e_2(t)$ appears at the receiving end of the system.

The output voltage may also be written in the form

$$e_2(t) = \int_{-\infty}^{\infty} [G(\omega) A(\omega)] e^{j[\omega(t-\tau)+\phi(\omega)-\theta(\omega)]}\, d\omega$$

The first term in brackets is the amplitude spectrum, and the second term, in the exponent, is the phase spectrum. Instead of grouping all the amplitude terms together and then all the phase terms together for the entire system, we may separate the amplitude and phase characteristics of the original signal from the amplitude and phase characteristics of the transmission path. Thus the output voltage may also be written as

$$e_2(t) = \int_{-\infty}^{\infty} [G(\omega) e^{j\phi(\omega)}] e^{j\omega t} [A(\omega) e^{-j[\omega\tau - \theta(\omega)]}]\, d\omega$$

where the first term in brackets is the amplitude and phase spectrum of the original signal, and the second term in brackets specifies the modifications of the amplitude and phase spectra of the signal due to the transmission properties of the system.

If the received signal is to be free from distortion introduced by the communication system, the second pair of brackets should equal unity. This requires that

$$A(\omega) = 1 \quad \text{and} \quad \theta(\omega) = -\omega\tau$$

In practice, as outlined in Sec. 48, we merely require $A(\omega) = k$ (a constant) since it may be desired to have the received signal amplified or attenuated to a value different from that of the original signal.

These results show that, for distortionless transmission, the amplitude spectrum of the transmission circuit should be a constant, independent of frequency, whereas the phase spectrum should have a phase shift proportional to frequency over a frequency range of $-\infty$ to ∞. When these conditions are fulfilled, the reproduced signal is delayed by an amount τ from that of the original signal.

If $A(\omega) = k$ and $\theta(\omega) = -\omega\tau$ are inserted into the above equation for $e_2(t)$, the received signal becomes

$$\begin{aligned} e_2(t) &= k \int_{-\infty}^{\infty} [G(\omega) e^{j\phi(\omega)}] e^{j\omega(t-\tau)}\, d\omega \\ &= k e_1(t - \tau) \end{aligned}$$

Hence the received signal will have its amplitude proportional to that of the original signal, and its reception will be delayed by a time interval $t - \tau$, but otherwise the original and reproduced signals will have the same wave form.

50. Distortion. In a broad sense, distortion is a change in wave form between an original wave and its re-created wave form. Several types of distortion are possible, at least some of which can be readily explained in terms of a simple sinusoid, expressible in terms of its amplitude, frequency, and phase.

In a linear passive system, *amplitude distortion* (sometimes called frequency distortion) is that form of distortion in which the relative magnitudes of the different frequency components are not constant, although the phase shift is linearly related to frequency.

Phase or *delay distortion* occurs when the phase angle of the system is not linearly related to frequency within a desired range, so that the time of transmission or the

delay time varies with frequency, even though the amplitude is constant for all frequencies within the specified range.

Amplitude and phase distortion may be present simultaneously in a system.

Nonlinear distortion is that form of distortion which occurs when the rms values of the voltage or current at the output of a system or portion of a system are not linearly proportional to the corresponding rms values of the input voltage or current. Nonlinear distortion occurs in nonlinear systems or elements, not in linear systems or elements.

Other types of distortion may exist, but those enumerated above are the most common.

51. Logarithmic Measurements. In communication engineering, considerable advantage is taken of the fact that an exponential or a logarithmic relation exists between variables. For example, many subjective sensations are estimated to be roughly proportional to logarithmic functions of the physical stimuli producing them. That is, if the physical stimulus is increased geometrically, the apparent subjective sensation produced by it appears to increase arithmetically.

The fact, however, that a subjective sensation appears to be approximately logarithmically related to the stimulus producing it is but a convenient, simple, and rough approximation in almost all cases. The correlation between subjective sensations and physical stimuli is exceedingly complicated, influenced by many environmental factors and techniques of measurement and evaluation. Nevertheless, a logarithmic relation is more nearly correct than a linear one and is a sufficiently good approximation in many cases to serve as a justification for the adoption of logarithmic units. They not only provide a useful (if approximate) relation between subjective sensations and physical stimuli but are also convenient in computation. A number of logarithmic units—particularly the *decibel*—have been found to represent such useful and convenient concepts that they have been improperly (and sometimes ridiculously) extended to fields in which they cannot apply.

The *bel*, named in honor of Alexander Graham Bell, is defined as the common logarithm of the ratio of two powers P_1 and P_2. Thus

$$\text{No. of bels} = N_b = \log_{10} (P_2/P_1)$$

If P_2 is greater than P_1, N_b will be positive (representing a gain in power); if $P_2 = P_1$, N_b will be zero; if P_2 is less than P_1, N_b will be negative and will represent a loss in power.

A smaller and more convenient unit, the decibel (db), has been established for engineering work. The magnitude of the *decibel* is one-tenth of a bel. The number of decibels required to express the change in power level from P_2 to P_1 is

$$\text{No. of db} = N_{db} = 10\ N_b = 10 \log_{10} (P_2/P_1)$$

A change of 1 db in acoustic power is just about the change in power level that can be detected.

Another logarithmic unit called the *neper* is defined as one-half the natural logarithm of the ratio of two powers. Thus

$$\text{No. of nepers} = N_n = \tfrac{1}{2} \log_e (P_2/P_1)$$

where e is the base of the natural system of logarithms and has the approximate value $e = 2.7182818285 \ldots$.

The relationships between the same power ratios when expressed in three systems are

$N_n = \frac{1}{2}(\log_e 10)N_b = 1.151292546 \ldots N_b$	or	1 bel = 1.151 nepers
$= \frac{1}{20}(\log_e 10)N_{db} = 0.1151292546 \ldots N_{db}$	or	1 db = 0.1151 neper
$N_{db} = 10N_b$	or	1 bel = 10 db
$= 20(\log_{10} e)N_n = 8.68588964 \ldots N_n$	or	1 neper = 8.686 db
$N_b = \frac{1}{10} N_{db}$	or	1 db = 0.1 bel
$= 2(\log_{10} e)N_n = 0.8685889 64 \ldots N_n$	or	1 neper = 0.8686 bel

These terms express only the ratio of power terms, and the number of units is not an expression for an absolute value of power. By assigning some value of power P_1

selected as the *reference level*, it is possible to ascertain an absolute value of power for any numerical expression in any of the three systems mentioned above. Values of 6, 10, and 100 mw have been used at the reference power level at different times and by different groups, so that the reference level should always be specified as well as the numerical units for a given change in power level. P_2/P_1 is always a numeric expressing an increase or decrease in power level and a given number of decibels or nepers cannot be used as an expression for an absolute value of power. When decibels are so used, some given value for P_2 or P_1 is implicit.

Because power in an electric circuit may be expressed in terms of either voltage or current and resistance, it is possible for any of the above expressions to take the general form (but not the exact value)

$$N = \log \frac{P_2}{P_1} = \log \frac{I_2^2 R_2}{I_1^2 R_1} = \log \left(\frac{E_2^2/R_2}{E_1^2/R_1}\right)$$

When a change of power level is expressed in terms of resistance and either current or voltage, it is generally necessary to express or specify the values of R_1 and R_2 with which I_1 and I_2 or E_1 and E_2 are associated. Only in the special case for which $R_2 = R_1$ does neglect of the values of the resistance introduce no error. It is common, if regrettable, practice, however, to overlook the values of the resistances when specifying power-level ratios in terms of voltages or currents.

The *volume unit* (*vu*) is defined as ten times the common logarithm of the ratio P_2/P_1, where P_1 is 1 mw (0.001 watt). If P_2 is expressed in watts,

$$\text{No. of vu} = N_{\text{vu}} = 10 \log_{10} (P_2/0.001) = 30 \log_{10} P_2$$

Since the reference level is specified in the definition of the volume unit, this term can be properly used to express either a change in power level or an absolute value of power.

The *phon* is a unit of loudness level. The level of a sound in phons is numerically equal to the intensity level (in decibels) of a pure 1,000-cps tone, which is judged by the listener to be of equivalent loudness. In establishing the unit for the phon, zero reference level is taken as the average limit of audibility at 1,000 cps; this has been standardized as 10^{-16} watt per sq cm, or a sound-level intensity of 0.000204 dyne per cm. Since the phon is a unit representing a subjective sensation, its value is directly affected by the average characteristics of the normal human ear. The number of phons representing a given sound-intensity level is, in general, different for each frequency and varies in a manner depending upon the sensitivity of the normal human ear, as shown in Fig. 10.

References

Analytical Approach to Communication Engineering:

Bennett, W. R.: Response of a Linear Rectifier to Signal and Noise, *J. Acoust. Soc. Am.*, **15** (3), 165, January, 1944.

Blackman, N. M.: The Demodulation of a Frequency Modulated Carrier and Random Noise by a Discriminator, *J. Applied Phys.*, **20** (10), 976, October, 1949.

Fano, R. M.: "The Transmission of Information," *Technical Report*, 65, M.I.T. Research Laboratory of Electronics, Mar. 17, 1949; *Technical Report*, 149, Feb. 6, 1950.

Knudtzon, N. H.: "Experimental Study of Statistical Characteristics of Filtered Random Noise," *Technical Report*, 115, M.I.T. Research Laboratory of Electronics, July 15, 1949.

Lee, Y. W.: "Application of Statistical Methods to Communication Problems," *Technical Report*, 181, M.I.T. Research Laboratory of Electronics, Sept. 1, 1950.

———, T. P. Cheatham, Jr., and J. B. Wiesner: "The Application of Correlation Functions in the Detection of Small Signals in Noise," *Technical Report*, 141, M.I.T. Research Laboratory of Electronics, Oct. 13, 1949.

——— and C. A. Strutt: "Statistical Prediction of Noise," *Technical Report*, 129, M.I.T. Research Laboratory of Electronics, July 12, 1949.

Middleton, D.: Some General Results in the Theory of Noise through Non-linear Devices, *Quart. Applied Math.*, **5** (4), 445, January, 1948.

———: The Response of Biased Saturated Linear and Quadrature Rectifiers to Random Noise, *J. Applied Phys.*, **17** (10), 778, October, 1946.

Information Theory:

Bode, H. W., and C. E. Shannon: A Simplified Derivation of Linear Least Square Smoothing and Prediction Theory, *Proc. IRE*, **38** (4), 417, April, 1950.
Gabor, D.: Theory of Communication, *J. Inst. Elec. Engrs. (London)*, **93**, Part III (26), 429, 1946.
Hartley, R. V. L.: Transmission of Information, *Bell System Tech. J.*, **7** (3), 535, July, 1928.
Jackson, W.: Information Theory, *Nature*, **167**, 20, 1951.
Kretzmer, E. R.: "Interference Characteristics of Pulse-time Modulation," *Technical Report*, 92, M.I.T. Research Laboratory of Electronics, May 27, 1949.
Lawson, J. L., and G. E. Uhlenbeck: "Threshold Signals," McGraw-Hill, 1950.
Middleton, D.: Rectification of a Sinusoidal Modulated Carrier in the Presence of Noise, *Proc. IRE*, **36** (12), 1467, December, 1948.
———: The Spectrum of Frequency Modulated Waves after Reception in Random Noise, Part I, *Quart. Applied Math.*, **7**, 129, 1949. Appendix, Part II, *Quart. Applied Math.*, **8**, 59, 1950.
Nyquist, H.: Certain Factors Affecting Telegraph Transmission Theory, *Bell System Tech. J.*, **20** (2), 119, April, 1941.
———: Certain Topics in Telegraph Transmission Theory, *Trans. AIEE*, **47**, 617, April, 1928.
Rice, S. O.: Mathematical Analysis of Random Noise, *Bell System Tech. J.*, **23** (3), 282, July, 1944; **24** (4), 46, January, 1945.
———: Communication in the Presence of Noise—Probability of Error for Two Encoding Schemes, *Bell System Tech. J.*, **29** (1), 60, January, 1950.
Shannon, C. E.: Communication in the Presence of Noise, *Proc. IRE*, **37** (1), 10, January, 1949.
———: A Mathematical Theory of Communication, *Bell System Tech. J.*, **27**, 379, July, 1948; **27**, 623, October, 1948.
Stumpers, F. L.: "A Bibliography of Information Theory—Communication Theory -Cybernetics," M.I.T. Research Laboratory of Electronics, Feb. 2, 1953.
"Symposium on Applications of Autocorrelation Analysis to Physical Problems," Woods Hole, Mass., June 13–14, 1949, Office of Naval Research, Department of the Navy.
"Symposium on Information Theory—1954," M.I.T., Sept. 15–17, 1954; *Trans. Prof. Group Information Theory, IRE.*
Tuller, W.: Theoretical Limitations on the Rate of Transmission of Information, *Proc. IRE*, **37** (5), 468, May, 1949.
Wiesner, J. B.: Statistical Theory of Communication, *Proc. Natl. Electron Conf.*, 1950.

References on Analytical Approach:

Black, H. S.: "Modulation Theory," Van Nostrand, 1953.
Cuccia, C. L.: "Harmonics, Sidebands, and Transients in Communication Engineering," McGraw-Hill, 1952.
Gabor, D.: Theory of Communication, *J. Inst. Elec. Engrs. (London)*, **93**, Part III (26), 429, November, 1946.
Goldman, S.: "Frequency Analysis, Modulation, and Noise," McGraw-Hill, 1948.
Guillemin, E. A.: "Communication Networks," Vol. 1, 1933; Vol. 2, Wiley, 1935.
Hershberger, W. D.: "Principles of Communication Systems," Prentice-Hall, 1955.
Jackson, W.: Information Theory, *Nature*, **167**, 20, 1951.
Root, W. L., and T. S. Pitcher: On the Fourier Series Expansion of Random Functions, *Annals of Math. Studies*, **26** (2), 313, June, 1955.
Shannon, C. E., and W. Weaver: "The Mathematical Theory of Communication," University of Illinois, 1949.

References on Logarithmic Units:

Chinn, H. A., D. K. Gannett, and R. M. Morris: A New Standard Volume Indicator and Reference Level, *Proc. IRE*, **28** (1), 1, January, 1940.
Gray, C. H., and W. H. Martin: Master Reference System for Telephone Transmission, *Bell System Tech. J.*, **8** (3), 536, July, 1929.
Hilliard, J. K.: Definition of Standard Reference System, *Electronics*, November, 1931.
Martin, W. H.: Decibel—The Name for the Transmission Unit, *Bell System Tech. J.*, **8** (1), 1, January, 1929.

PSYCHOPHYSICAL RESPONSES AND THE COMMUNICATION SYSTEM

52. Man and His Environment. Man's knowledge of the universe is the combined result of impressions gained from physical stimuli to one or more of his senses, and the thought processes used to evaluate these impressions.

Man's visual response is to electromagnetic energy lying between about 400 and 750 millimicrons ($m\mu$) in wavelength corresponding to spectral colors from violet through blue, green, yellow, orange, to deep red. The eye is not uniformly sensitive in this range but has maximum sensitivity at about 554 $m\mu$, corresponding to the green portion of the spectrum. The color sensitivity of the eye also depends upon the intensity of the radiation and is different for very low light levels than for levels normally encountered in daylight.

Man's two eyes, separated by the interpupillary distance (about 2.5 in. or 10 cm), see slightly different images, which are combined in the brain to yield a single interrelated impression. The sensations produced by two different images enable man to perceive depth or a third dimension through stereoscopic vision. Stereoscopic visual effects are not usually employed in electrical communication.

If the visual images are seen in color, the eye is able to make good distinction between portions of the image which are in different colors. If the images seen are in monochrome, the original color of an object makes its contribution to the visual image only through varying shades of monochrome and portions of the images are distinguished by their relative densities.

Man's aural response is to compressional disturbances in an elastic medium—almost always air—covering the frequency range of from about 20 to about 20,000 cps. The sensitivity of the normal human ear is greatest to frequencies of about 1,000 to 2,000 cps, although the ability to perceive the full a-f range depends upon the age of the person and upon any hearing defects he may have.

Man's two ears, on opposite sides of the head, separated by a distance of about 6 or 7 in., hear sounds from a given source at slightly different times. When the two sounds are heard and combined in the brain, man is able to perceive depth or a third dimension through stereophonic hearing; he is able to determine reasonably well, therefore, the point in space from which the sound emanates.

53. Role of Sensory Perceptions in Communication. The type or character of the signal to be transmitted through a communication system depends upon the kind of intelligence to be conveyed. It also depends upon the sense which is to be stimulated at the receiver. Whatever sense is to be stimulated, appropriate energy-converting devices, transducers, are required at each end of the system. For example, in broadcasting, microphones are necessary to convert the acoustic energy to electrical energy at the transmitter and loud-speakers are required at the receiver to convert the transmitted electrical energy into acoustic energy.

The satisfactory design of energy-converting devices involves three factors:

1. Physical means for converting sensation-stimulating energy into electrical form and vice versa.
2. Means for determining the performance of the energy-converting devices in terms of physical energy as well as in terms of the sensations produced by energy converters. The criterion of distortionless transmission, already given, may often be employed to guide design. It may be technically impossible or economically unsound to construct distortionless system, but by having knowledge of the conditions for distortionless communication, the engineer can recognize the limitations beyond which practice cannot advance.
3. Knowledge of certain physiological characteristics of human beings, such as the relations between sensations or subjective responses and the physical stimuli producing such responses. It is particularly important to know, in as quantitative a manner as possible, the significant characteristics of speech, hearing, and vision, since such knowledge is of direct aid in the design of significant components of the communication system.

If the system is to be designed and built for use by normal persons, it is desirable to know, quantitatively, the normal human responses as determined from a statistically significant group of persons. The data obtained from a study of the subjective sensations and characteristics of a statistically adequate group may be used in the design of communication equipment for any type of service required by normal persons; but the obtaining of such information is not easy.

54. Evaluating Sense Perceptions. How is the communications engineer to determine the performance of the communication system in terms of measurements that are, in some way or another, related to sense perceptions? The desired objective is clear enough. What we would like to have is a set of instruments whose readings or indications of physical stimuli would correspond with the subjective evaluations of such stimuli when judged by a normal person occupying the same position as the instrument. In acoustics, for example, it would be useful to have an instrument to

measure sound pressure with readings proportional to the human subjective evaluation of the sound pressure.

There is a good deal of difficulty in providing such instruments, and only a beginning has been made. The real problem is that of determining the subjective estimates or evaluations of the physical phenomena evoking sensations. No direct measurements can be made; all correlations between the physical stimuli and estimates of the subjective effects produced must be done indirectly with groups of people making the necessary evaluations. Those selected for such tests must have normal responses, and this poses very real problems. A large number of persons must be employed and they must be properly conditioned. Finally, the kind of tests to be given, the conditions under which stimuli are produced and sensations evaluated, and the significance of the results obtained are all important problems in correlating the physical and the psychophysical phenomena.

Although, in making such psychophysical tests, each observer may obtain a different result, usually it will be found that, under identical conditions of measurement, observations by a large number of persons tend to cluster around a relatively small range of values, especially for persons whose responses are regarded as being "normal." Substantial deviations, on the part of a single observer or small number of observers, from the results obtained by a large number of persons are a very effective means for determining deviations from normal, especially if results from the "deviationists" tend to be inconsistent.

Data obtained from large numbers of persons by techniques such as these have sometimes been adopted as standard, after proper examination and evaluation of experimental technique and consistency of results. The standard audibility and visibility curves, for example, have been obtained in this way.

55. Hearing. The human ear is not responsive to pressure variations in a compressible medium over an infinite range of sound intensities. Some sounds are of such low energy level that they cannot be heard; other sounds may have such a high energy level that they produce pain or physical damage rather than the sensation of sound. The loudness of sounds is recognized as such only between the audibility threshold of hearing (the minimum level at which sounds can be heard) and the maximum or tolerance level of hearing (at which pain is produced).

The ear is, likewise, not equally sensitive to sounds of different frequencies. Sounds below about 15 to 30 cps are not perceived as sounds, whereas vibrations beyond about 15 or 20 kc are beyond the upper frequency range of audibility. At the extremes of hearing, the threshold of audibility merges with the threshold of pain.

For normal human ears, maximum sensitivity occurs for sounds from about 800 to 2,500 cps. Toward the ends of the a-f range, the ear is not as sensitive to low-intensity sounds as in the middle of the range. The thresholds of audibility and of pain tend to merge so that there is a smaller range of sound levels the normal ear recognizes at the frequency extremes than in the middle of the audible band.

Responses vary from person to person; they vary with the age of the individual and may even depend upon the audio environment prior to test, and persons having hearing defects have different responses than those with normal hearing.

The two ears are separated by a distance of about 6 or 7 in., and each receives a slightly different signal from the same sound source. Some estimate of the direction of a single sound source can be determined by turning the head so that each ear receives sound of equal intensity and with the same time delay.

56. Audibility Threshold. The threshold of hearing a specified sound is the minimum effective pressure of that sound which is capable of producing an auditory sensation in a noise-free environment in a specified fraction of all observations or tests. It is usually specified in decibels above a standard sound-pressure intensity of 0.0002 dyne per sq cm, which represents the threshold of hearing a pure 1,000-cps tone by persons having the most acute hearing.

The American Standard threshold of audibility is determined for a person with acute hearing listening to pure tones. Measurements are made in an anechoic

chamber with the observer facing the source of sound, which is at a distance greater than 1 m from him. The sound pressure is measured at a point where the observer's head will be located. Audibility-threshold curves determined in this way are shown as curve *A* of Fig. 9. Curve *B* of the same diagram represents the threshold of audibility when the sound is supplied by earphones and the sound pressure is determined at the entrance to the ear canal.

The two different methods of measuring audibility threshold disclose two significant differences. Curve *A*, for binaural hearing, shows that the threshold of audibility is about 10 db lower when two ears are involved than for one-ear, or monaural, hearing,

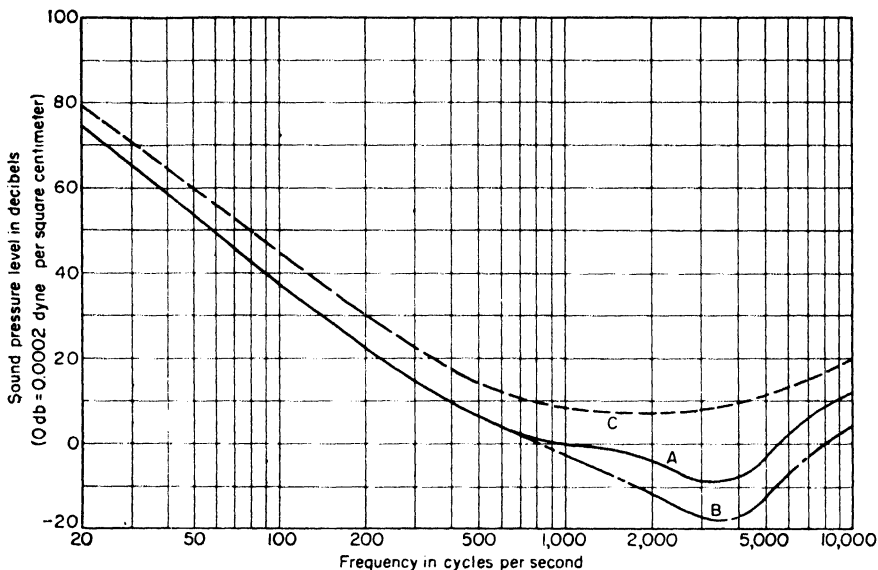



FIG. 9. Thresholds of audibility as determined by three different procedures. Curve *A* is for binaural hearing, with sound produced by single sound source in front of listener. Curve *B* is for binaural hearing with sound produced by number of randomly distributed sound sources. Curve *C* is for monaural hearing, with sound conveyed to listener by headset.

represented in curve *C*. The response for binaural hearing also has greater irregularities above about 800 cps. These are attributed to acoustical resonances in the outer ear canal, as well as to diffraction of sound around the observer's head.

A frequency of 20,000 cps is usually regarded as the upper frequency limit of audibility, but it varies with the sound level; it is different for different persons and is higher for young persons than for middle-aged or old people.

57. Tolerance Threshold. When sound is conveyed by earphones, listeners begin to experience discomfort when pure tones reach levels greater than 110 db and a tickling sensation when the tones reach levels of 130 db, reference level of 0 db being taken at a sound pressure of 0.0002 dyne per sq cm. Pain is experienced at sound levels of 140 db. Except at the extremes of the a-f range, these levels are not greatly dependent upon frequency. Persons customarily exposed to intense sounds have a higher tolerance threshold in each of the three categories than are indicated by the figures given above.

When the sound is noise instead of a pure tone, essentially the same sensations are aroused at the intensity levels given. It is not necessary, however, for all frequency components of wide-band spectrum noise to have the indicated level to produce the

stated sensation; when the energy in a portion of the spectrum reaches 110, 130, or 140 db, the sensations of discomfort, tickling, and pain, respectively, are usually encountered.

58. Loudness and Loudness Levels. In acoustics it is much more customary to measure the loudness level than to measure the loudness, which is the magnitude of the sensation. Loudness level is the intensity level of an equally loud reference tone, usually 1,000 cps. Loudness level is an experimentally determined quantity and is measured in decibels above a specified reference level. The reference level is 10^{-16} watt per sq cm, corresponding to sound pressure of 0.000204 dyne per sq cm and 0 db sound intensity.

For normal human ears, the average characteristics relating frequency and loudness level are shown in Fig. 10. The curves were obtained by alternately listening to a

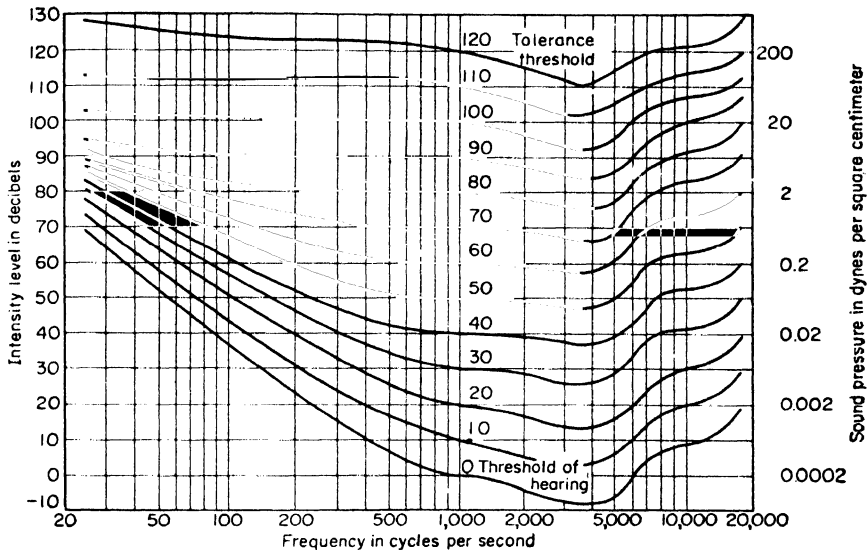



Fig. 10. Contours of equal loudness for normal human ears. Numbers on the curves indicate the loudness levels.

sound of arbitrary frequency and intensity and comparing it with a tone of 1,000 cycles whose intensity was varied until the two sounds were judged to be equally loud. The curves of Fig. 10 show the relative insensitiveness of normal ears at low intensity levels for high and especially for low frequencies. The maximum range of intensity levels to which the ear responds is about 140 db, representing a power ratio of 10^{14} to 1 from tolerance level to threshold of audibility. The numbers on the different curves represent the estimated loudness levels in phons. The loudness level is a subjective evaluation in which any tone is compared with a tone of 1,000 cps and variable intensity. Since the comparisons are made at 1,000 cps, the loudness levels in phons and the intensity level in decibels are the same at a frequency of 1,000 cycles. At 30 cps a loudness level of 10 phons occurs when the intensity level is 70 db.

59. Minimum Perceptible Changes in Sound Pressure. A change in sound-pressure level of about 1 db can be detected for tones between 50 and 10,000 cps, provided the level of the tone is greater than 50 db above that tone's threshold value. If the sound-pressure level is less than 40 db above the threshold value, changes of from 1 to 3 db are required to be perceptible. Under unusually favorable conditions levels as small as 0.3 db can be detected by normal ears in the middle of the a-f range.

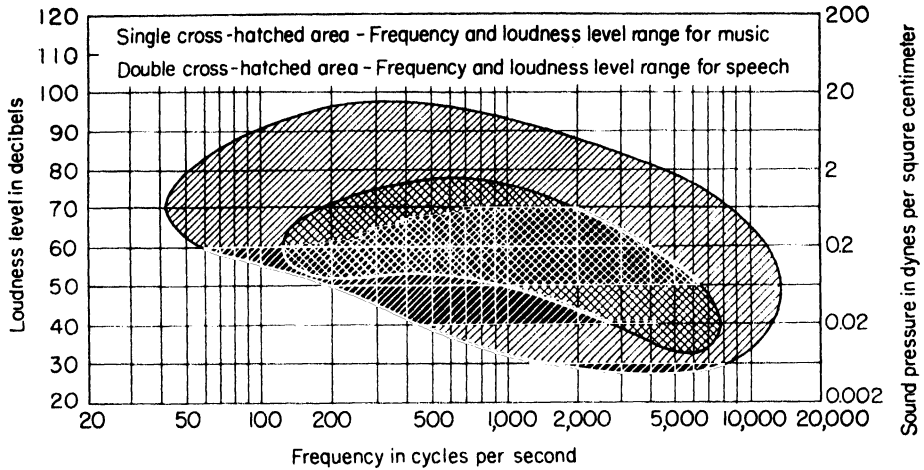



FIG. 11. Frequency and power ranges for the reproduction of speech and music.

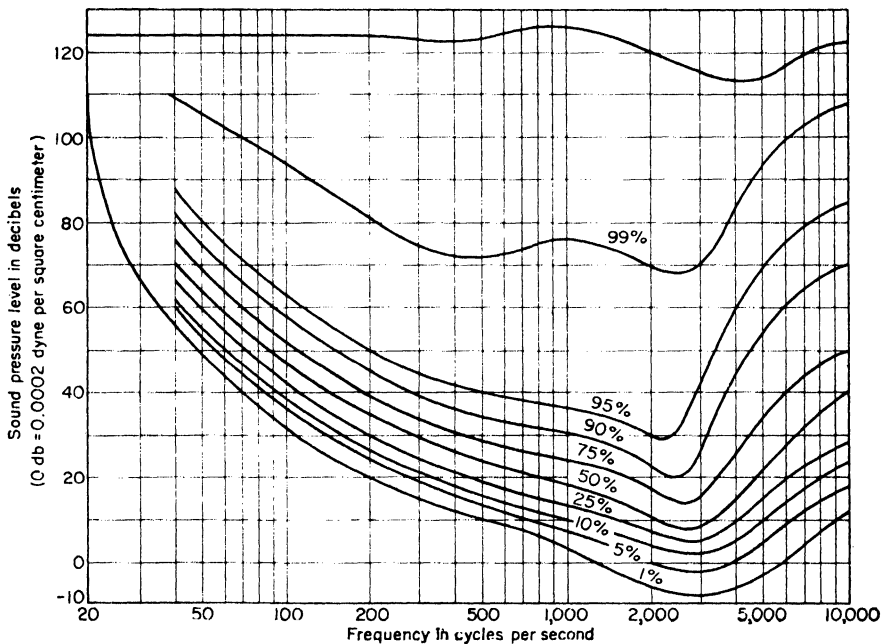



FIG. 12. Thresholds of hearing for statistically large sample of American population. Numbers on the curves indicate the percentage of persons tested whose threshold of audibility was lower than that represented by the curve marked; *i.e.*, their hearing was more acute. Top curve represents threshold of feeling.

60. Minimum Perceptible Changes in Frequency. The ear can detect changes in frequency of about 0.3 of 1 per cent above 1,000 cps if the pressure levels are greater than 40 db above threshold value, or about 3 cps below 1,000 cps. For low levels of pressure, the minimum perceptible frequency change may be many times this set of values, especially at low frequencies.

61. Frequency and Power Ranges for Practical Aural Communication. From the contour curves for equal loudness for human ears, it may be determined that distortionless transmission of sounds by electrical means requires the transmission of frequencies of from 20 to 20,000 cps and the transmission of power ranges of as much as 10^{14} to 1. Few systems respond fully without distortion over this range, and none of them is capable of operating effectively over the full range of power ratios to which the ear is normally responsive.

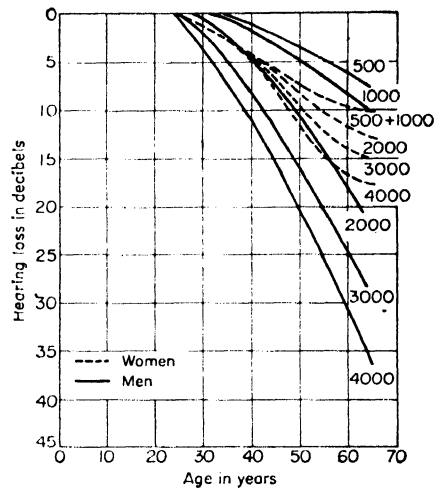
Fortunately it is not necessary to reproduce the complete range of frequencies or loudness levels to obtain satisfactory, understandable, or enjoyable reproduction of speech and music. Estimates vary as to the frequency range and intensity levels that are desirable, particularly since economic factors must be taken into account, for extension of frequency range and intensity levels invariably is accompanied by increased cost. Furthermore, some types of audio service are more stringent in their demands than others.

FIG. 13. Hearing impairment as a function of age. Dashed curves represent data for women, solid curves for men. Numbers on the curves indicate the frequency at which observations were made.

The frequency and power ranges required for the satisfactory reproduction of speech and music are shown in Fig. 11. According to these data, little improvement in the reproduction of music is obtained by extending the frequency range below 40 cps or above 15 kc or by extending the intensity levels below 28 db or above 97 db. For persons of normal hearing, practically no gain justified by the increased cost will be achieved in transmitting speech at frequencies below 130 or above 5,000 cps, or by transmitting power levels outside those from 35 to 78 db. Indeed, intelligible speech can be reproduced with a frequency band of from 250 to 2,800 cps and for the reproduction of music, many broadcast receivers are so operated that their frequency band is limited to between about 100 and 3,000 cps. The highest-quality sound systems, on the other hand, are usually designed to reproduce frequencies from 20 to 20,000 cps.

Deviations from normal hearing occur in a significant fraction of the population and depend on age and sex as well as individual characteristics. They must be considered in economic communication systems. Figures 12 and 13 illustrate these deviations.

62. Sensation-Stimulus Relations. From the loudness-level curves of Fig. 10, it is possible to derive curves in which the loudness level is plotted against the sound pressure, as in Fig. 14. Such a relation shows directly the subjective evaluation of loudness level as a function of stimulus, within the framework of the experimental technique employed. If the Weber-Fechner law were rigorously true, the curves of Fig. 14 would be straight lines.

The psychophysical response of the human ear is not a linear function of the amplitude of the stimulus, if a sufficient range of the latter is taken into account. Measurements by Stevens and Newman show that, although the stimulus-sensation response

is linear for small values of sound-pressure intensity, amplitude distortion is produced within the ear itself for loud amplitudes of sound.

63. Frequency and Pitch. Frequency is a physical quantity measured with physical apparatus and specifies the number of oscillatory vibrations per unit of time.

Pitch, on the other hand, is an aspect of auditory sensation in terms of which sound may be ordered on a scale, such as a musical scale. The subjective quantity of pitch is primarily a function of the physical frequency of a sound, but not entirely so, for the pitch of pure tones is also dependent upon the sound intensity. The subjective determination of pitch is determined experimentally in terms of the physical property of frequency. An oscillator of known frequency and sound intensity is fed to a

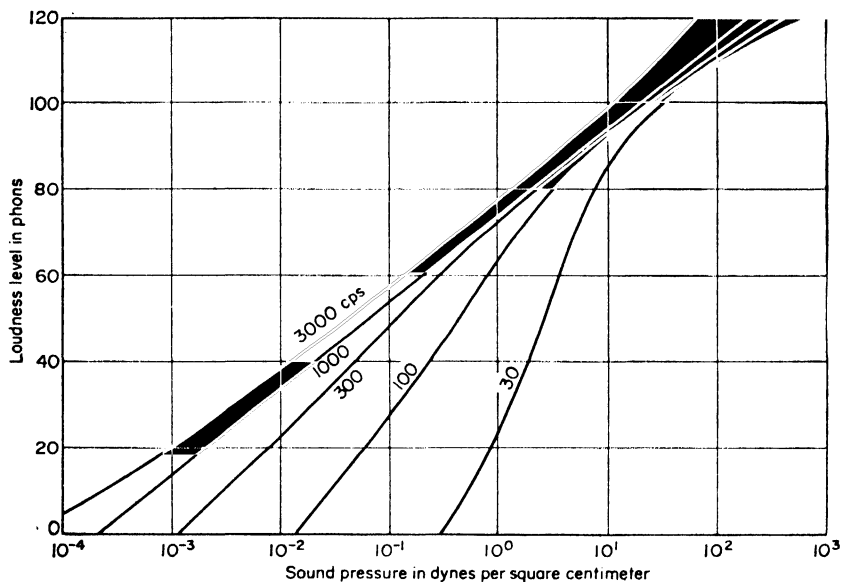



FIG. 14. Relative loudness (psychophysical response) plotted against sound pressure (physical stimulus) for five different audio frequencies.

headset or loud-speaker. Another headset or loud-speaker is fed from another oscillator whose frequency is adjustable. For a range of frequencies of the first oscillator, the observer adjusts the frequency of the second oscillator until the pitch of the sound produced by both oscillators has some specified relationship. With such an experimental technique, a scale of pitch can be related to the frequency. Pitch is measured in units of the *mel*, and a reference pitch of 1,000 mels was chosen as the pitch of a pure tone of 1,000 cps having a sound-pressure level of 60 db.

The relation between subjective pitch, in mels, and frequency, in cps, throughout the entire a-f range of 20 to 20,000 cps is shown in Fig. 15. It is clear that pitch, as determined by the experimental technique outlined above, is by no means linearly related to the frequency of the sound.

64. Phase Shift. Conclusive evidence has not been obtained to indicate that the ear can evaluate phase shift in the components of a complex sound. The response of the ear to complex sounds appears to depend upon the magnitude of the harmonic content rather than upon the wave shape, which, for a specified amplitude-frequency spectrum, is determined by the phase shift of the harmonics compared with that of the fundamental wave. For this reason the phase-shift characteristics of a-f portions of communication systems can be, and frequently are, neglected.

Although the ear itself cannot distinguish the phase relations of the harmonics

of a complex sound, subsidiary (although usually minor) matters are observed as a result of phase shift. In a complex wave, the phase relations of the harmonics may influence the subjective sensations of a complex tone. Provided the fundamental frequency is less than 100 cps, a shift in phase of a harmonic may alter the subjective loudness of the harmonic; it may also produce a noticeable difference in the quality of the complex tone.

65. Masking Effects. A sound impressed upon the ear reduces the ability of the ear to hear other sounds and may even completely mask their presence. A sound is said to be masked by another when the intensity of the second sound is increased

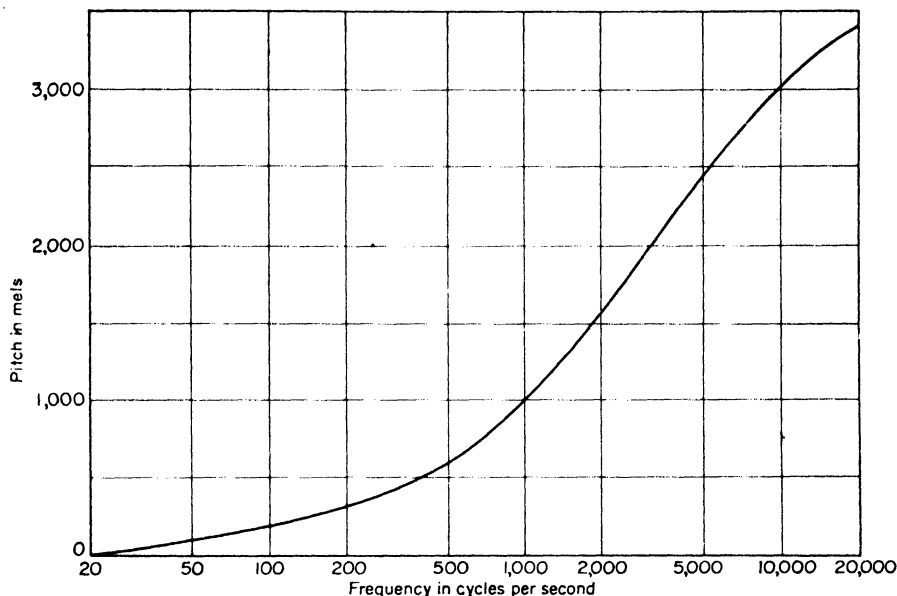



FIG. 15. Tests of the evaluation of pitch indicate that a nonlinear relation exists between the observer's estimate of the pitch of a sound and its frequency. The two are taken to be equal at 1,000 cps.

to such a level that the first can no longer be heard. Masking is defined as the number of decibels by which the listener's threshold of audibility for hearing a given sound is increased by the presence of a second or masking sound. Suppose, for example, that, in the absence of noise or any other sound, the threshold of audibility for a given sound is 3 db. The presence of a second sound will make the first one inaudible until the level of the first has been raised to a higher level, say 36 db. Thus, the second tone produces a masking effect of 36 − 3, or 33 db.

When one pure tone is masked by another pure tone, the greatest amount of masking occurs when the two tones are nearly alike in frequency. For low-intensity levels of the masking tone, the masking effect is usually small when the frequency of the masking tone is double that of the original tone. As the intensity of the masking tone is increased, the masking effect of the interfering tone remains high so that all the tones of high frequency must be increased in intensity to fairly high levels before they can be heard. The effect of masking a pure tone by another pure tone is greater when the masking or interfering tone is above the tone being masked than when below it in frequency, when the two tones are heard simultaneously by one or both ears.

If the original and masking tones are introduced separately to each of the two ears the interfering effects are not the same as when introduced simultaneously into one

70. Speech Clipping. Limitation of the peak amplitude in speech is often employed in communication systems to improve the intelligibility of the received signal for a given amount of power generated and modulated at the transmitter.

The average power content in speech wave forms is appreciably less than that of a sine wave of the same amplitude value as the peak of the speech wave. The percentage of modulation is determined in terms of peak amplitudes, and for this reason, the power in the modulated wave or in the side bands of a transmitter modulated 100 per cent by speech wave forms will be much less than if the same transmitter were completely modulated by a sine wave.

If the magnitude of the speech wave form is now increased so that its average power is the same as that of the sine wave and if, furthermore, the amplified speech wave is clipped so that its maximum values do not exceed the amplitude of the sine wave, the speech will be able to modulate the transmitter 100 per cent. Of course the wave form of the clipped speech will differ from that of unclipped speech wave forms; the clipped speech will lose some of its naturalness, but the degradation for moderate amounts of clipping is not great so far as concerns articulation. Moreover, once the system is properly adjusted, the maximum amplitude of the speech is maintained at such a fixed value as to make overmodulation impossible.

In the process of clipping, certain undesirable high-frequency distortion components are generated, and these should be removed by filters, prior to modulation. Suitable filters should have relatively little attenuation for frequencies below 2,500 cps and high attenuation for frequencies above about 3,000 to 3,500 cps.

With 6 to 12 db of clipping the change in quality of speech is small, but four to sixteen times as much speech power can be transmitted as without clipping. Clipping of as much as 25 db can be employed without too much loss of intelligibility.

References

Physical Basis of Sound:

Beranek, L. L.: "Acoustic Measurements," Wiley, 1949.
———: "Acoustics," McGraw-Hill, 1954.
Crandall, I.: "Vibrating Systems and Sound," Van Nostrand, 1926.
Fletcher, H.: Physical Characteristics of Speech and Music, *Bell System Tech. J.*, **10** (3), 349, July, 1931.
———: "Speech and Hearing in Communication," Van Nostrand, 1953.
Grimsehl, A.: "A Textbook of Physics," Blackie, 1933.
Miller, D. C.: "The Science of Musical Sounds," Macmillan, 1937.
Miller, G. A.: "Language and Communication," McGraw-Hill, 1951.
Morse, P. M.: "Vibrations and Sound," 2d ed., McGraw-Hill, 1948.
Potter, R. K., G. A. Kopp, and H. C. Green: "Visible Speech," Van Nostrand, 1947.
Watson, F. R.: "Sound," Wiley, 1945.
Wood, A. B.: "Sound," Macmillan, 1944.

Psychophysical Aspects of Sound:

Bartholemew, W. T.: "Acoustics and Music," Prentice-Hall, 1942.
Beasely, W. C.: Characteristic Distribution of Impaired Hearing in the Population of the United States, *J. Acoust. Soc. Am.*, **12** (1), 114, July, 1940.
Best, E. H., and N. B. Taylor: "Physiological Basis of Medical Practice," Sec. 9, Williams & Wilkins, 1945.
Castner, T. G., and C. W. Carter: Developments in the Application of Articulation Testing, *Bell System Tech. J.*, **12** (3), 347, July, 1933.
Crandall, I. B.: Sounds of Speech, *Bell System Tech. J.*, **4** (4), 586, October, 1925.
———: Speech, Dynamical Study of Vowel Sounds, *Bell System Tech. J.*, **6** (1), 100, January, 1927.
——— and D. MacKenzie: Analysis of Energy Distribution in Speech, *Bell System Tech. J.*, **1** (1), 116, July, 1922.
Fletcher, H.: The Nature of Speech and Its Interpretation, *Bell System Tech. J.*, **1** (1), 129, July, 1922.
———: Physical Measurements of Audition and Their Bearing on the Theory of Hearing, *Bell System Tech. J.*, **2** (4), 145, October, 1923.
———: Pitch, Loudness, and Quality of Musical Tones, *Am. J. Phys.*, **14** (4), 215, August-September, 1946.
———: Loudness, Pitch and Timbre of Physical Stimulus, *J. Acoust. Soc. Am.*, **6** (2) 59, October, 1934.
——— and W. S. Munson: Loudness; Its Definition, Measurement, and Calculation, *J. Acoust. Soc. Am.*, **5** (2), 182, October, 1933; *Bell System Tech. J.*, **7** (4), 377, October, 1933.
——— and J. C. Steinberg: Articulation Testing Methods, *Bell System Tech. J.*, **8** (4), 806, October, 1929.
——— and ———: Loudness of a Complex Sound, *Phys. Rev.*, **24** (9), 306, September, 1924.

French, N. R., and J. C. Steinberg: Factors Governing Intelligibility of Speech Sounds, *J. Acoust. Soc. Am.*, **19** (1), 90, January, 1947.
Howell, W. H.: "A Textbook of Physiology," 10th ed., Saunders, 1927.
Ingalls, A. H., C. H. G. Gray, and R. T. Jenkins: A Voice and Ear for Telephone Measurements, *Bell System Tech. J.*, **11** (2), 293, April, 1932.
Knudsen, V. O.: Sensibility of the Ear to Small Differences of Intensity and Frequency, *Phys. Rev.*, **2** (1), 82, January, 1923.
Licklider, J. C. R.: Effect of Amplitude on the Intelligibility of Speech, *J. Acoust. Soc. Am.*, **18** (2), 429, October, 1946.
Polyak, S. I., G. McHugh, and D. K. Judd: "The Human Ear in Anatomical Transparencies," McJenna, 1946.
Riesz, R. R.: An Artificial Larynx, *J. Acoust. Soc. Am.*, **1,** Part 1 (2), 273, January, 1930.
———: Differential Intensity Sensitivity of the Ear for Pure Tones, *Phys. Rev.*, May, 1928, p. 867.
Robinson, D. W., and R. S. Dodson: Threshold of Hearing and Equal Loudness for Pure Tones and Loudness Function, *J. Acoust. Soc. Am.*, **29** (12), 1284, December, 1957.
Sacia, C. F., and C. J. Beck: Power of Fundamental Speech Sounds, *Bell System Tech. J.*, **5** (3), 393, July, 1926.
Shower, E. G., and R. Biddulph: Differential Pitch Sensitivity of the Ear, *J. Acoust. Soc. Am.*, **3,** Part 1 (2), 275, October, 1931.
Sivian, L. J.: Speech Power and Its Measurement, *Bell System Tech. J.*, **8** (4), 646, October, 1929.
———, H. K. Dunn, and S. D. White: Absolute Amplitudes and Spectra of Musical Instruments and Orchestras, *J. Acoust. Soc. Am.*, **2** (3), 330, January, 1931.
Snow, W. B.: Audible Frequency Range of Music, Speech, and Noise, *J. Acoust. Soc. Am.*, **3,** Part 1 (1), 155, July, 1931.
———, H. C. Montgomery, and M. B. Gardner: Results of World's Fair Hearing Tests, *Bell System Tech. J.*, **19** (4), 533, October, 1940.
Steinberg, J. C.: Application of Sound Measuring Instruments to the Study of Phonetic Problems, *J. Acoust. Soc. Am.*, **6** (1), 16, July, 1934.
———: Loudness of a Sound and Its Physical Stimulus, *Phys. Rev.*, October, 1925, p. 507.
———: Articulation Testing and Methods, *Bell System Tech. J.*, **8** (4), 806, October, 1929.
Stevens, S. S.: Calculating the Loudness of Noise, *Science*, **123** (3199), 676, Apr. 20, 1956.
——— and H. Davis: "Hearing, Its Psychology and Physiology," Wiley, 1938.
———: "Handbook of Experimental Psychology," Wiley, 1953.
Stewart, J. Q.: An Electrical Analogue of the Vocal Organs, *Nature*, September, 1922, p. 311.

Hearing and Acoustics:

Beranek, L. L.: "Acoustic Measurements," Wiley, 1949.
———: "Acoustics," McGraw-Hill, 1954.
———: Design of Speech Communication Systems, *Proc. IRE*, **36** (9), 880, September, 1947.
———, A. P. G. Peterson, J. L. Marshall, and A. L. Cudworth: Calculation and Measurement of Loudness of Sounds, *J. Acoust. Soc. Am.*, **23** (5), 261, May, 1951.
———, W. H. Radford, J. A. Kessler, and J. B. Wiesner: Speech Reinforcement System Evaluation, *Proc. IRE*, **39** (11), 1401, November, 1951.
——— and H. W. Rudmouse: Sound Control in Airplanes, *J. Acoust. Soc. Am.*, **19** (3), 357, March, 1947.
Bolt, R. H., and A. D. MacDonald: Theory of Speech Masking by Reverberation, *J. Acoust. Soc. Am.*, **21** (11), 557, 1949.
Davis, A. H.: "Modern Acoustics," Macmillan, 1927.
——— and C. J. Morreau: "Reduction of Noise in Buildings," H. M. Stationery Office, 1939.
Egan, J. P., and H. W. Hake: On the Masking Pattern of Simple Auditory Stimulus, *J. Acoust. Soc. Am.*, **22,** 622, September, 1950.
Fletcher, H.: Auditory Patterns, *Revs. Modern Phys.*, **12** (1), 47, January, 1940.
French, N. R., and J. C. Steinberg: Factors Governing the Intelligibility of Speech Sounds, *J. Acoust. Soc. Am.*, **19** (1), 90, January, 1947.
Hirsch, I. J.: "The Measurement of Hearing," McGraw-Hill, 1952.
Hughes, L. E. C.: "Elements of Engineering Acoustics," Benn, 1953.
Institute of Radio Engineers: Proceedings, the Transactions of Professional Groups on Audio Systems and Broadcast Transmission Systems.
Knudsen, V. O.: "Architectural Acoustics," Wiley, 1940.
———: "Modern Acoustics and Culture," University of California Press, 1937.
——— and C. M. Harris: "Acoustical Design in Architecture," Wiley, 1950.
Kryter, K. D.: The Effect of Noise on Man, *J. Speech and Hearing Disorders*, Monthly Supplement 1, September, 1950.
McLachlan, W. N.: "The New Acoustics," Oxford, 1936.
Motion Picture Arts and Sciences, Academy: "Motion Picture Sound Engineering," Van Nostrand, 1938.
Olson, H. F.: "Elements of Acoustical Engineering," Van Nostrand, 1940.
Rosenblith, W. A.: "Relations of Hearing Loss to Noise Exposure," American Standards Association, New York.
——— and K. N. Stevens: "Handbook of Acoustic Noise Control," Vol. 2, Noise and Man, WADC Technical Report 52-104, June, 1953, Office of Technical Services, Department of Commerce, Washington, D.C.
Sabine, P. E.: "Collected Papers on Acoustics," Harvard University Press, 1922.

Schafer, T. H.: Frequency Selectivity of the Ear as Determined by Masking Experiments, *J. Acoust. Soc. Am.*, **22** (7), 490, July, 1950.

71. Vision and the Communication System. The introduction of picture and television communication systems has made it imperative that the communication engineer have some acquaintance with the properties of light as well as the visual characteristics of the human eye, for the requirements of satisfactory visual communication systems are much more stringent than those to be fulfilled for a satisfactory aural system. Only the most elementary matters of vision and visual responses can be mentioned here; for more details reference should be made to specialized works on optics, vision, psychology, or "human engineering."

72. Light and Vision. In a physical sense the term "light" refers to electromagnetic radiation of such wavelength as will produce a visual sensation in the normal human eye. The term light is sometimes also used to refer to wavelengths shorter than or longer than those of the visible spectrum. Vision, on the other hand, depends upon the ability of the eye to transform light stimulus into impulses that are transmitted to the brain through nerve fibers. The undistorted perception of contrast and color, of shade, depth, motion, direction, and most voluntary thought processes depend upon the response of the eye to light. Thus, vision is a subjective response to a physical stimulus.

Frequently the term "light" is used to describe either the radiant energy capable of producing a visual sensation, or the sensation that is produced by radiant energy. Physicists usually use the term in its physical or objective sense. On the other hand, physiologists and psychologists usually use the term in its subjective sense.

From the physical point of view, light is a form of electromagnetic radiation similar to radio waves except for their wavelength or frequency. Thus, except for those factors depending upon wavelength or frequency, the comments given here on light apply equally well to radio waves.

In free space light travels with a velocity of 2.99796×10^{10} centimeters per second (186,000 miles per second). Light is a transverse wave motion consisting of electric and magnetic vectors at right angles to each other and to the direction of propagation. It may be polarized, reflected, absorbed, refracted, absorbed, or transmitted in various degrees by materials upon which it falls.

Table 3. Approximate Wavelength and Frequency Limits of Colors in the Visible Spectrum

Spectrum color	Frequency limits, cps	Wavelength limits, mμ
Infrared	Less than 4.0×10^{14}	Greater than 750
Red	$4.0\text{–}4.8 \times 10^{14}$	750–630
Orange	$4.8\text{–}5.0 \times 10^{14}$	630–600
Yellow	$5.0\text{–}5.2 \times 10^{14}$	600–580
Green	$5.2\text{–}5.9 \times 10^{14}$	580–510
Blue	$5.9\text{–}6.5 \times 10^{14}$	510–460
Violet	$6.5\text{–}7.5 \times 10^{14}$	460–400
Ultraviolet	Greater than 7.5×10^{14}	Less than 400

Selective reflection (or transmission) and absorption account for the colored appearances of objects seen by the eye. Thus, an object which absorbs to a high degree a portion of the visual spectrum will present to the eye the color sensation corresponding to white light minus the portion absorbed. That is, the only portion affecting the eye is the portion reflected (or transmitted) to the eye.

The human eye is a highly specialized organ, remarkably (but not uniformly) sensitive to electromagnetic energy between the approximate limits of 400 and 750 millimicrons. Light, impinging upon the eye and focused by the lens, produces an image on the retina whose form, change of position, intensity, and spectral energy

istribution are subjectively interpreted in terms of shape, motion, brightness, color, ..d other attributes of objects in the physical world.

The eye can distinguish details of the image and differentiate distant objects from nearby ones, in part, by the relative sharpness of detail of their images. The relative distances and positions are inferred from the angles between the lines of sight from the eyes to the objects in question, and by the relative angles subtended by objects of known size. The eye receives sensations of color from light of different wavelengths but is not able to analyze a given color sensation into its spectral components.

The eye has the ability to admit more or less light to the retina by varying the pupillary aperture, and thereby can adapt human vision to a wide range of ambient light intensity or brightness.

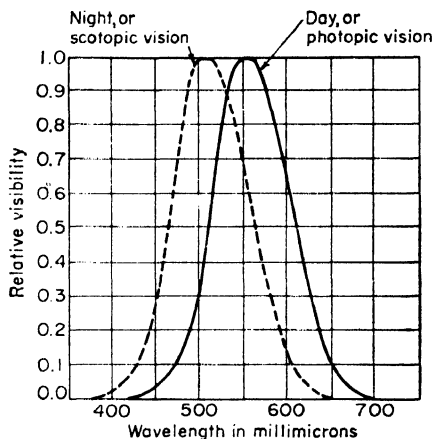
From a point of fixation straight ahead, the field of vision of the human eye extends more than 90° outward, 70° downward, 60° inward, and 50° upward. The two eyes see slightly different visual images because of their interpupillary separation. The slightly different views seen by the different eyes give rise to stereoscopy, by which the relative size and distance of objects are also determined.

73. Standard Visibility Function. The human eye is not equally sensitive to all wavelengths it is able to detect. For normal levels of illumination, the eye responds to all wavelengths between about 400 and 750 millimicrons; it is least sensitive to wavelengths at the extremes of this band and most sensitive to an intermediate or mean value in the yellow-green portion of the spectrum. Different eyes (even in the same individual) may have different spectral sensitivity or visibility functions.

FIG. 17. Visual sensitivity of the normal human eye for ordinary day vision and for vision at very low levels of intensity.

If the energy of the visual spectrum is passed through a prism and narrow slit so that monochromatic radiation of any selected value can be impressed on the eye and if the intensity of the radiation can be varied at will, it will be found that maximum spectral sensitivity occurs at 554 millimicrons for the "normal human observer." The relative sensitivity of the normal human eye to monochromatic radiation of other wavelengths (for light of normal intensity) is shown in Fig. 17. If the intensity of the light is substantially reduced, the spectral sensitivity curve shifts to shorter wavelengths, becoming maximum at 515 millimicrons. The spectral sensitivity, or standard visibility, curve shows the relative brightness of radiant energy for equal energy at each wavelength in the visible spectrum.

74. Contrast Sensitivity. If an object is to stand out from a background having the same color, contrast in brightness must exist between the two when both are illuminated by the same radiation. The minimum detectable difference in brightness between two large, contiguous surfaces is used as a test for contrast and contrast sensitivity. The surfaces used should subtend an angle of several degrees at the eye, and the light intensity of the two fields is adjusted until the observer notes a just perceptible difference between the two. The ratio of the difference in brightness of the two fields to the mean brightness is called the *contrast*. *Contrast sensitivity* is the reciprocal of the contrast as defined above.

Contrast and contrast sensitivity depend upon the surround of the illuminated areas. If the surround is illuminated, much greater values of contrast sensitivity can be determined than if the surround is dark. If the surround is illuminated, contrast sensitivity increases linearly with the logarithm of the background illumination,

whereas, if the surround is dark, saturation effects occur and a small range of contrast is observed, as shown in Fig. 18.

75. Resolving Power. When there is an appreciable amount of contrast between an object and its background, the eye has the ability of distinguishing between objects having small angular separation, and can also distinguish fine detail.

The degree to which the eye is able to distinguish objects separated by a small angle is called resolving power and depends upon the kind and shape of the object, the degree of contrast between it and its surround, the spectral distribution of the energy of illumination, the criterion used to determine or evaluate perception of the object, and other factors. Depending upon the experimental conditions of the test, this property lies in the range of about 0.1 to 1 minute of visual angle; that is, under the most favorable circumstances, objects can be seen separately if, at the observer's eye, they subtend an angle of 0.1 to 1 minute of arc.

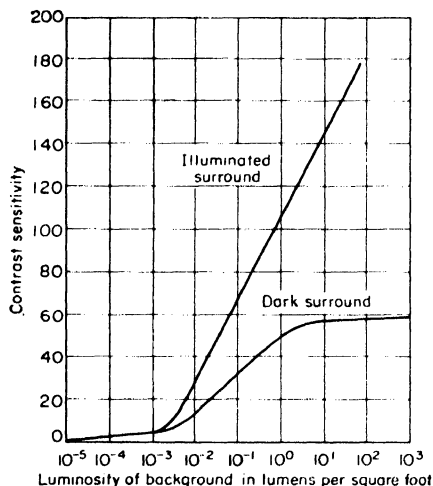



FIG. 18. Contrast sensitivity of the human eye at various levels of background luminosity for illuminated and dark surrounds.

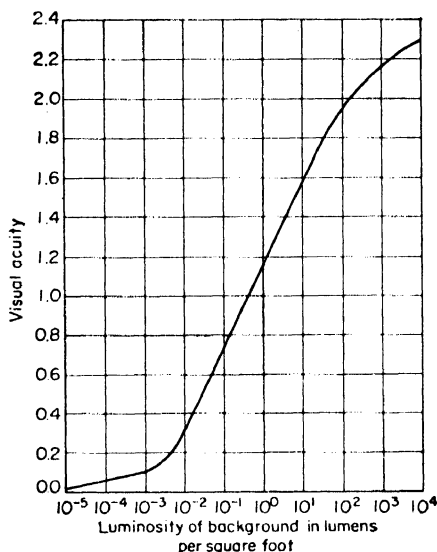



FIG. 19. Visual acuity as a function of the luminosity of the background for normal human eyes.

76. Visual Acuity. Whereas resolving power is related to the minimum detectable visual angle, visual acuity is the ability of the eye to distinguish fine detail. Sometimes visual acuity is expressed as the reciprocal of the angle, in minutes, which the smallest detail in a test object subtends at the eye; thus visual acuity becomes equivalent to the reciprocal of the resolving power. Acuity may also be expressed as either (1) the ratio of the distance at which a given line of letters on a test chart can be seen by the observer being tested, to the distance which an observer with normal vision would see it, or (2) a visual rating (expressed in per cent) related to the size of characters in lines of letters of different size. The first method is employed with the Snellen test chart; the second with the test chart of the American Medical Association. As shown in Fig. 19, visual acuity increases with the amount of illumination, or the brightness of the image.

77. Flicker and Persistence of Vision. A visual image may be fully perceived by a viewer in about 0.01 sec after it has been projected onto a screen. The visual impression does not immediately vanish upon being removed from the screen suddenly, but persists as an afterimage for about 0.05 sec. This phenomenon, called the per-

sistence of vision, is employed in motion pictures and in television to create the illusion of motion by exhibiting, in rapid succession, a series of images that are similar to one another but individually different in those portions of the frame in which motion occurs. The effects of persistence of vision depend upon the luminosity of the subject, and below a certain interruption rate (which depends upon luminosity) the persistence of vision vanishes and flicker is observed. Flicker is usually annoying and objectionable.

If 20 or more frames or images are viewed per second from a sequence of frames which are alike except where motion has occurred between the making of one frame and the next, and if the successive images from each frame can be reconstructed before the visual impression of the preceding one has been lost, the sensation produced is one of apparent continuous motion, and the images are reasonably free from flicker. As shown in Fig. 20, the frequency of flicker is not constant but depends upon the field luminosity.

78. Color Vision. The eye is incapable of analyzing a complex stimulus into its spectral components; in this respect there is a fundamental difference between the eye and the ear. Because the eye cannot analyze radiation into its spectral components, an infinite number of stimuli will, in general, evoke a given hue sensation.

It is found that a color stimulus can be matched visually by proper combination of three color attributes. Thus, three numbers (which may be chosen in different ways) are generally necessary, and are always sufficient, to specify the sensation-evoking characteristics of a given radiation.

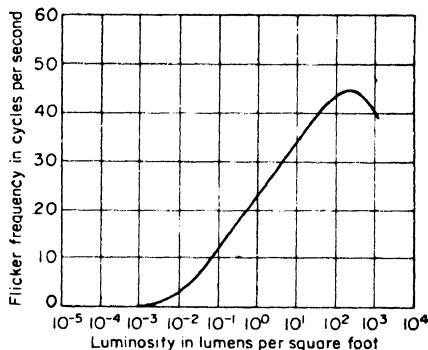
FIG. 20. Critical flicker frequency for human eyes.

A color match may be produced by means of a pure spectral color to which varying amounts of neutral (gray) light have been added. In this method, color may be specified in terms of the physical quantities of (1) dominant wavelength, (2) purity, and (3) luminosity, which correspond roughly to the subjective sensations of (1) hue, (2) saturation, and (3) brightness, respectively.

A color match may also be produced by mixing three stimuli, called *primaries*. To produce a match of the proper brightness, the intensity of the three primaries may need to be varied as well as their relative amounts, which determine the hue of the mixture. The primaries for achieving visual color matches are usually taken to be red, green, and blue, but they may be other hues—or even monochromatic radiation—and may be arbitrarily chosen so long as some amounts of all three components are needed to produce a subjective color match. Color specification in terms of amounts of the three primaries required to produce a match does not correspond to any commonly accepted sensation attributes.

The spectral energy curves of three primaries established by international agreement in 1931 are shown in Fig. 21. The primaries are selected so that equal amounts of all three are required to produce white light. A convenient simplification has been achieved by selecting one of the primaries to coincide with the visibility function of the normal human observer.

79. Hue, Saturation, and Brightness. Every color sensation has three qualities which we may call (1) hue, (2) degree of saturation, and (3) brightness. These qualities are related to the physical attributes of (1) dominant wavelength, (2) purity, and (3) luminosity, and form the basis of color television systems of matching color sensations.

Hue is that attribute of color by virtue of which it differs characteristically from gray of the same brightness, and in respect to which it falls into classes we may

designate as red, orange, yellow, green, blue, or violet, or intermediates of these. Hue may be expressed as a spectral wavelength; in physical measurements it is designated as the dominant wavelength.

The degree of *saturation* of a color refers to the absence of neutral or gray light. Pure spectral colors without admixture of gray (white) light are said to be saturated; they are less saturated in proportion to the amount of white light added to effect a color match. Saturation is a term that describes the absence of white light mixed with the hue. Pale or pastel shades of hue are less saturated than the vivid shades; pink is white with some red and is but slightly saturated.

The *brightness* of a color depends upon the intensity of the energy reaching the eye. Brightness is a characteristic by means of which colors are located on a scale ranging from light (white) to dark (black). Saturation and brightness are somewhat related, for saturation refers to the degree by which a color departs from gray or neutral hue of the same brightness.

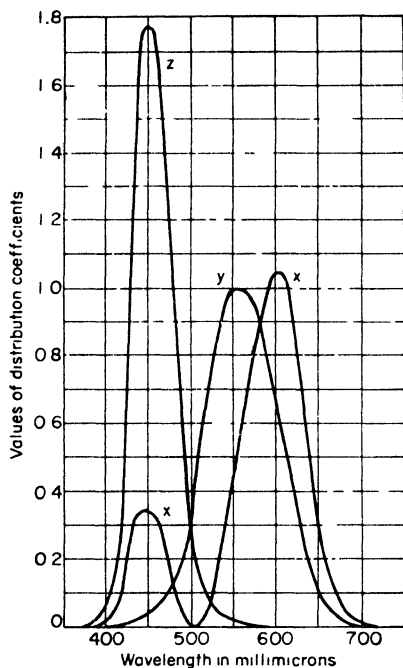



FIG. 21. Distribution coefficients for C.E.I. primaries selected for colorimetric work. The curve marked y is the same as the visual sensitivity of the eye for high light intensities.

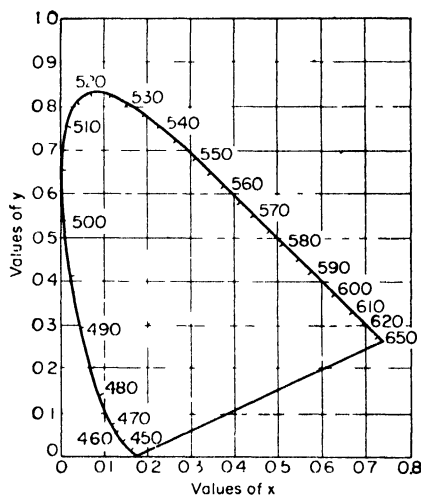



FIG. 22. Chromaticity diagram or color triangle. The horseshoe-shaped curve shows the locus of all spectrum colors, the wavelength in millimicrons of the spectral colors being designated by the numbers on the curves. White light is represented at the center of the diagram (x = 0.31; y = 0.31).

Hue and saturation, together, are sometimes called *chroma*. Thus, the chromaticity characteristics tell everything about the color except its intensity or brightness.

To completely specify a color sensation, it is necessary to give the hue (the wavelength of the dominant hue), the quantity of white light mixed with it (or the degree of saturation), and the intensity of the color (as determined by its brightness or brilliance). Two colors are said to be visually alike when they are in agreement in all of these three respects.

80. Representation of Colors. A basic tool in the study and specification of color is the *chromaticity diagram*. Chromaticity is that characteristic of color representing hue and saturation together; it describes two of the three properties of color, omitting brightness.

The most commonly used chromaticity diagram is the kind of *color triangle* shown in Fig. 22 and based on the color mixture curves or primaries of Fig. 21. The color mixture curves show the amounts of the three primaries, X, Y, and Z, that are required to match unit energy at each wavelength in the visible spectrum. The diagram of Fig. 22 may be regarded as a standardized color map for the system of colorimetry used by the International Commission of Illumination. The primaries used in the National Television Systems Committee signal specifications are given in this system of color measurement and appear as points in the diagram of Fig. 22.

81. Colorimetric Basis of Picture Transmission. Because any color may be matched by proper amounts of three primary colors, any color image may be resolved into three partial images each of which represents the contribution of one primary-color component to the total colored image. Thus the red, green, and blue partial images represent, respectively, the amount and spatial distribution of the red, green, and blue hues of the original colored images; except that the partial images analyze the original view on a color basis, they have the same size and amount of detail as the original colored image. Each partial color image may be considered to be a monochrome visual image, on a color basis, of the original.

Each of the three partial images may now be analyzed into picture elements by a suitable scanning process, converted into magnitude-time functions of electrical phenomena, and synthesized at the receiving end into another partial image. So far as each partial image is concerned, the process is the same as the electrical transmission of a monochrome picture. Now, however, three partial images must be transmitted—one for each primary color—and each partial image reconstructed at the receiving end of the system must control or produce light of the appropriate primary stimulus. When this is done and the three reconstructed partial images are superimposed, in register, the original color image may be reconstructed.

If a single still view is to be transmitted in color, the time of transmission is usually not of major importance, and a fairly narrow band width can be used to transmit colored images by electrical means. On the other hand, if colored pictures are to portray motion, each set of three partial images for any given frame of the colored image must be scanned, its electrical wave form must be transmitted, and the new image must be reconstructed in something like 0.01 sec if the effects of flicker are to be minimized.

It is possible to transmit all of one partial image at one time, and to follow this by the transmission of all of the remaining partial images. But it is also possible to transmit one element or one line of the red-color image, followed by one element or line of the green-color image, and then one element or line of the blue-color image, in repetitive fashion. Details of the various systems in current use are covered in other chapters of this handbook.

References

Physical Aspects of Light and Optics:

Bouma, P. J.: "Physical Aspects of Color," Philips Gloeilampenfabrikens, 1947.
Evans, R. M.: "An Introduction to Color," Wiley, 1948.
Hardy, A. C., and F. H. Perrin: "The Principles of Optics," McGraw-Hill, 1932.
Houston, R. A.: "A Treatise on Light," Longmans, 1930.
"I.E.S. Lighting Handbook," Illuminating Engineering Society, New York, 1947.
Luneberg, R. K.: "Mathematical Analysis of Binocular Vision," Princeton University Press, 1947.
Monk, G. S.: "Light—Principles and Experiments," McGraw-Hill, 1937.
Moon, P.: "Scientific Basis of Illuminating Engineering," McGraw-Hill, 1936.
Sears, F. W.: "Optics," 3d ed., Addison-Wesley, 1949.
Southall, J. P. C.: "Introduction to Physical Optics," Oxford, 1937.

Psychophysical Aspects of Light and Vision:

Adams, D.: "Dark Adaptation; A Review of Literature," Oxford, 1929.
Bartlett, N. R., and A. L. Sweet: Visibility on Cathode Ray Tube Screens, Signals on a P7 Screen Exposed for Different Intervals, *J. Optical Soc. Am.*, **39** (6), 470, June, 1949.
———, S. B. Williams, and R. M. Hanes: Visibility on Cathode Ray Tube Screens; The Effect of Size and Shape of the Pip, *J. Optical Soc. Am.*, **39** (6), 463, June, 1949.
Best, C. H., and N. B. Taylor: "Physiological Basis of Medical Practice," Williams & Wilkins, 1945.
Bouman, M. A.: Peripheral Contrast Thresholds of the Human Eye, *J. Optical Soc. Am.*, **40** (12), 825, December, 1950.

Bryan, G. M.: Physical and Photochemical Basis of Visual Resolving Power—Part II—Visual Acuity and the Photochemistry of the Retina, *J. Optical Soc. Am.*, **34** (12), 718, December, 1944.
Cady, F. E., and H. B. Dates: "Illumination Engineering," Wiley, 1928.
Carmichael, L., and F. W. Dearborn: "Reading and Visual Fatigue," Houghton Mifflin, 1947.
Detweiler, S. R.: Comparative Anatomical Studies of the Eye, with Especial Reference to the Photoreceptors, *J. Optical Soc. Am.*, **30** (2), 42, February, 1940.
Ferree, C. E., and G. Rand: "Studies in Physiological Optics," Vols. I and II, Wilmer Ophthalmological Institute, 1934.
Fulton, J. F., P. M. Hoff, and H. T. Perkins: "Bibliography of Visual Literature, 1938–1944," Office of Scientific Research and Development, 1945.
Hanes, R. M., and S. B. Williams: Visibility on Cathode Ray Tube Screens: The Effects of Light Adaption, *J. Optical Soc. Am.*, **38** (4), 363, April, 1948.
Higgins, G. C., and K. Stultz: Variation of Visual Acuity with Various Test-objective Orientations and Viewing Conditions, *J. Optical Soc. Am.*, **40** (3), 135, March, 1950.
Hunt, R. W. G.: The Effects of Daylight and Tungsten Light-adaption on Color Perception, *J. Optical Soc. Am.*, **40** (6), 362, June, 1950.
Jahn, T. L.: Visual Critical Flicker Frequency as a Function of Intensity, *J. Optical Soc. Am.*, **36** (2), 76, February, 1946.
———: Brightness Discrimination and Visual Acuity as Functions of Intensity, *J. Optical Soc. Am.*, **36** (2), 83, February, 1946.
Klüver, H.: "Visual Mechanism," Jaques Cattell Press, 1942.
Knoll, H. A., R. Tonsey, and E. O. Hulbert: Visual Thresholds of Steady Sources of Light in Fields of Brightness from Dark to Daylight, *J. Optical Soc. Am.*, **36** (8), 480, August, 1946.
Kronfeld, P. C., G. McHugh, and S. L. Polyak: "The Human Eye in Anatomical Transparencies," Bausch and Lomb, 1943.
Luckiesh, M.: "Light, Vision and Seeing," Van Nostrand, 1944.
——— and F. K. Moss: The Human Seeing Machine, *J. Franklin Inst.*, **215** (6), 629, June, 1933.
——— and F. K. Moss: "The Science of Seeing," Van Nostrand, 1937.
——— and R. K. Moss: "Seeing—A Partnership of Lighting and Vision," Williams & Wilkins, 1931.
Lythgoe, R. J.: "Measurement of Visual Acuity," Oxford, 1932.
Moon, P.: "Scientific Basis of Illuminating Engineering," McGraw-Hill, 1936.
——— and D. E. Spencer: Visual Data Applied to Lighting Design, *J. Optical Soc. Am.*, **34** (10), 605, October, 1944.
Murray, H. D., and D. A. Spencer: "Color in Theory and Practice," Vol. 2, "General Theory," American Photographic Publishing Company, 1939.
Newhall, S. M., and J. G. Brennan: "Comparative List of Color Terms," Intersociety Color Council, 1949.
Otero, J. M., L. Plaza, and F. Saloveri: "Absolute Thresholds and Night Myopia," *J. Optical Soc. Am.*, **39** (2), 167, February, 1949.
Pirenne, M. H.: "Vision and the Eye," Pilot Press, 1948.
Schouten, J. F., and L. S. Ornstein: Measurements on Direct and Indirect Adaptation by Means of a Binocular Method, *J. Optical Soc. Am.*, **29** (4), 168, April, 1939.
Sinden, R. H.: Some Consequences of the Additive of Luminosities, *J. Optical Soc. Am.*, **40** (10), 647, October, 1950.
Sweet, A. L., and N. R. Bartlett: Visibility on Cathode Ray Tube Screens; Signals on a P7 Screen at Different Intervals after Excitation, *J. Optical Soc. Am.*, **38** (4), 329, April, 1948.
Von Schelling, H.: A Method for Calculating the Effect of Filters on Color Vision, *J. Optical Soc. Am.*, **40** (7), 419, July, 1950.
Wald, G., and D. R. Griffin: The Change in Refractive Power of the Human Eye in Dim and Bright Light, *J. Optical Soc. Am.*, **37** (5), 321, May, 1947.

ELECTROMAGNETISM AND THE STRUCTURE OF MATTER

82. Properties of Materials. Communication engineering is concerned with analytical methods of expressing and representing intelligence, particularly in terms of the human responses of the senses—especially hearing and seeing. The basic concepts of this phase of communication have already been outlined for any type of communication system. Electrical-communication systems are concerned with the utilization of electromagnetism as the vehicle by which information, originating at one part of the system, may be conveyed to another portion of the system. In radio-communication systems, conveyance of information through the use of electromagnetic radiation is a primary concern. Electrical-communication engineering thus rests squarely upon the fundamentals of electromagnetism, which is the subject of the remainder of this chapter.

Engineering applications of electricity involve the use of material bodies, not only to give form, embodiment, and mechanical strength to the necessary structures, but also to make possible the creation of electric and magnetic fields of the desired

magnitude and spatial distribution, together with the desired and effective utilization and storage of electromagnetic energy.

In the design of communication apparatus, the physical, thermal, optical, acoustical, chemical, and especially the electromagnetic properties of materials play important roles in making possible the desired performances. A vast amount of investigation has shown that the properties of materials depend, in large measure, upon their structure in terms of the elementary particles of nature, and upon the processing to which they are subject. It is not yet possible to predict or predetermine the properties of materials in terms of their composition and structure, but considerable progress is being made in this field, which has been referred to as molecular engineering.

The subject of the properties of materials, as determined by their structure in terms of molecules, atoms, and the more elementary particles of nature, is such a vast and extensive one that specialized treatises should be consulted for further information. Only the barest outline of a few factors upon which the electromagnetic properties of materials depend can be given here. Already significant progress has been made in relating the conductive, dielectric, and magnetic properties of materials with their composition and structure, and the time may come when it may be possible to specify the electromagnetic properties of materials and have suitable substances made to order with the desired characteristics. Indeed, the rapid progress that has been made in semiconductor devices is, in large measure, attributable to a thorough understanding of the relation between the properties of materials and their composition and structure.

83. Structure of Matter. Many experimental observations, particularly within the last half century, have conclusively demonstrated the atomic nature of electrical and magnetic phenomena, and the close dependence of electromagnetic effects upon the ultimate structure of matter.

So far as the engineer is concerned, most if not all of the phenomena with which he deals can be explained by the simple atom model composed of a compact stable central core or nucleus made up of relatively dense particles called *neutrons* and *protons*. Around the core electrons revolve in elliptical shells or orbits at distances large compared with the size of the nucleus. The electrons spin on their own axes.

At least 27 fundamental particles were recognized by the end of 1957, but most of these have very short lives (small fractions of a second) and do not enter into electromagnetic phenomena ordinarily encountered by the communications engineer. The stable particles of physics are shown in Table 4.

The electromagnetic properties of matter are largely associated with these electrons, and the communications engineer need not concern himself too much with the central core or nucleus.

The modern view of electricity regards an electric current (at least the conduction and convection components of the total current) as the flow of negative charges in one direction through a given area, plus the flow of positive charges (when these are free to move) through the same area in the opposite direction. In most cases the negative charges accounting for current are due to the motion of electrons, whose properties are given in Table 5. By common agreement, the positive sense of the direction of current is taken as that in which positive charges tend to move or (what is the same thing) opposite that to which negative charges move or tend to move. This convention was adopted before the true nature of the electric current was known, and since (in most cases) the electric current is constituted by the flow of electrons, the anomalous and awkward situation arises that the direction of flow of electricity is frequently in a direction opposite to the flow of charged particles that produce the current! For engineering-design problems, it makes little difference what the true direction is, but when this is important, as in the case of rectifiers, for example, care should be taken to specify whether the direction of flow is that of the electrons or of positive charges.

The motion of electric charges is produced by the application of electrical forces to material substances. The application of such an electrical force tends to dissoci-

ate particles of the substance into charged particles whose motion is produced by the electrical force and constitutes a current. The charged particles may be electrons, removed from atoms, or electrically charged atoms, called *ions*. A substance such as any of the metals, carbon, alkaline or saline solutions, or ionized gases, in which relatively large drift of electrons or other charged particles is produced by relatively small externally applied electrical force, or in which the charges may travel freely in distributing themselves within the substance, is called a *conductor*.

Table 4. Properties of Elementary Particles

Name	Symbol	Relative mass	Relative charge	Spin	Principal characteristics
Photon	γ	0	0	1	A quantum of electromagnetic energy, of magnitude $h\nu$, where h is Planck's constant, and ν is the frequency associated with the photon
Neutrino	ν	0	0	½	Existence postulated to account for continuous energy distribution of β particle, and required to conserve angular momentum in β decay processes; has negligible reaction with matter
Antineutrino	$\bar{\nu}$	0	0	½	Emission is hypothesized to accompany radioactive decay by positive emission or electron capture, or to conserve energy when proton changes to a neutron
Electron	e^-	1	−1	½	Fundamental stable particle, making up structure of atoms; arrangement of electrons in atomic orbits accounts for chemical and electrical properties of atoms
Positron	e^+	1	+1	½	Positive electron, formed in β decay of many radio nuclides, in pair production, and in other processes; secondary constituent of cosmic rays
Proton	p	1,835	+1	½	Equivalent to nucleus of hydrogen atom, and constituent of all atomic nuclei; accounts for charge of nuclei and about half of nuclear mass
Neutron	n	1,837	0	½	Neutral-charge, electron-proton combination; constituent of all nuclei having mass greater than 1; half-life, about 12 minutes

Table 5. Properties of the Electron

Rest mass of electron, m_0	9.1066×10^{-31} kg
Charge of electron, e	1.60203×10^{-19} coulomb
	1.60203×10^{-20} abcoulomb
	4.8025×10^{-10} statcoulomb
Ratio of charge to mass (e/m_0)	1.7592×10^{11} coulombs/kg
Size of electron (roughly)	1.9×10^{-15} m

A substance such as sulfur, glass, porcelain, mica, rubber, quartz, very pure water, many oils, and un-ionized gases, in which measurable amounts of current may be produced (without disrupting the material itself) only by the application of comparatively large electrical force, or in which charges are bound within certain small regions and are not free to distribute themselves readily within the substance, is called a *dielectric* or *insulator*.

There is no sharp dividing line between conductors and insulators. The classification of a material into conductors or insulators depends in part on its application. A substance that might be regarded as a good insulator for some applications might also be regarded as a fairly good conductor for other applications.

Substances intermediate between good conductors and good dielectrics are sometimes also called *semiconductors*, although within the past decade this term has come to designate a certain class of materials having rather specific properties in addition to poor conductivity.

Conductivity depends upon the ease with which charged particles, free to move, may be produced under the influence of an electrical force, and this in turn depends upon the atomic structure. Dielectric behavior, on the other hand, depends upon the degree to which charged particles are tightly bound within the atoms and molecules of the material, and again, these properties depend upon atomic and molecular structure.

Finally, the magnetic properties of a substance—and particularly its ferromagnetic properties—depend not alone upon the structure of individual atoms but also upon the manner in which atoms are grouped into regions, called *domains*, and the behavior of the various domains in the entire substance. Thus the crystalline structure of the ferromagnetic material is of considerable importance.

84. Electrical Conduction. Electrical conduction depends upon the availability of charged particles that are free to move through the circuit. The number and kind of charged particles depend upon the state of the conductor, *i.e.*, whether it is a solid, a liquid, or a gas.

In the conduction of electricity through gases, the normal gas atom is ionized by the electric field to which it is subjected. In this process one or more electrons are stripped from the outer shells or orbits. Under the influence of the electric field, the electrons are accelerated in one direction, and the remaining part of the atom—a positive ion—is accelerated in the opposite direction, but with considerably less mobility because of its greater mass.

In the conduction of electricity through liquids—electrolytic conduction—the charged particles are the positive and negative ions formed by electrolytic dissociation. If the liquid is monatomic, the positive ions are merely atoms of the liquid from which one or more electrons have been removed, whereas negative ions are atoms having an excess of their normal complement of electrons. If the molecules are composed of more than one kind of atom, they may dissociate into their atomic form but with electric charge, so that each dissociated particle contains electric charge but has mass essentially that of the atom rather than that of the electron. Both kinds of ions are active elements of conduction current. Unlike the situation in gaseous conduction, however, in electrolytic conduction the positive and negative ions have much more nearly the same mass and mobility, and therefore the total current is more evenly divided between the two kinds of ions of opposite charge.

In the conduction of electricity through a vacuum or by metallic conduction the positively charged particles are not free to move, and for such cases, the electric current is due entirely to the flow of electrons.

The flow of electricity through metals is due to the motion of free electrons within the metal. The crystals of solids are pictured as a three-dimensional lattice of atoms. The atoms are regarded as being essentially fixed in their positions, although they can vibrate about their positions of equilibrium. Metallic conductors for electrical use are formed from crystals in which the lattice structures are almost identical.

The metals are usually fairly dense and have a complicated atomic structure. The outer-shell electrons—the valence electrons—move within a complicated electric field composed of the positive field due to the protons in the nucleus, and the negative field of the extranuclear electrons. The outer-shell electrons are less firmly attached to the atom, not only because they are further from the field of the protons but also because the positive proton field is more or less neutralized by the negative field of the inner-orbit electrons.

The valence electrons are loosely connected to their respective atoms, and can be removed rather easily by an electric field and can thus be made free to travel between the atoms of the lattice. The good conducting properties of the metals are attributed to the fact that there are something like 10^{22} free electrons per cubic centimeter in metallic crystals.

At very low temperatures, the free electrons of metals are able to move readily under the influence of an electric field. At higher temperatures, the thermal agitation of the atoms of the metals introduces a certain impediment to free motion which is

manifest by an increase in electrical resistance. Metallic conduction is dependent upon the temperature as well as upon the crystal structure of the solid; it may also be appreciably influenced by the presence or absence of impurities which affect the homogeneity of the structure of the crystals of the conducting solid. The effect of impurities is much less in good conductors than in semiconductors, however.

85. Semiconduction. The term semiconductor has been used to designate a class of materials having conductivities intermediate between those of good conductors and good dielectrics. Semiconductors have negative coefficients of resistivity, their electrical properties are sensitive to impurities and methods of preparation, and the type and degree of atomic binding force differ from those of good conductors.

The high conductivity of metals commonly used in electrical engineering is the result of very weak bonds between the valence electrons and the atomic nuclei. In semiconductors, however, the valence electrons are an integral part of the structure of the substance; accordingly they are relatively fixed and fairly tightly bonded in their positions in the atom. For semiconductors made of ideal crystals in which the electrons are firmly bound within the crystal structure, no conduction would be possible and the material would behave as a dielectric. Some slight impurities always exist, or the crystal structure may not be perfectly symmetrical; in such cases a slight excess or deficiency of electrons may be available for conduction. The number of available electrons or the number of *holes* (absence or deficiency of electrons) is so small, compared with the number of atoms, that poor conductivity results.

When a few excess electrons are available, they can travel through the lattice structure of the crystal under the influence of an electric field, and this constitutes a conduction current. If the material has a dearth of electrons, it is said to possess "holes." A *hole*, therefore, may be considered as the absence of an electron, or it may be considered as equivalent to a positive charge. If a material containing holes is subjected to an electric field, electrons from adjacent atoms will progressively move to fill the hole. As the electrons move in one direction to fill the hole, the hole may be regarded as traveling in the opposite direction. Thus the hole can account for conduction just as an electron can, although the hole moves in a direction opposite to that of the electron.

Electrons and holes can be introduced into the lattice structures of metals by impurities, but a few electrons and holes can occur in the semiconducting metal itself. Semiconductors are found in the middle of the periodic table, where the valence of the material is 4. Materials of valence 3 or 5, when introduced as impurities in small amounts, increase conduction by injecting holes or electrons. When the crystal structure is exceedingly pure, the behavior of the semiconductor is almost like that of a dielectric or insulator. On the other hand, impurities to the extent of but 0.001 per cent are sufficient to make the material behave as a good conductor.

86. Dielectric Behavior. Dielectrics have no appreciable electrical conductivity. As a consequence of their nonconducting properties, they must be able to support electric strain when subject to the influence of an electric field. They should also return to the system the energy they absorbed when strain was created within the dielectric by an electric field. Materials composed of atoms from the central part of the periodic table tend to produce good dielectrics.

These atoms are neither strongly electropositive nor strongly electronegative; they usually form chemical bonds, in compound formation, by sharing outer-shell or valence electrons to complete their outer octet of electrons. Materials having such bonds—called *covalent bonds*—contain electrically balanced molecules in which ions are not produced, and therefore electrical conduction does not occur.

Good dielectrics are frequently composed of molecules made up of slightly electropositive and slightly electronegative atoms held together through covalent bonding. Although the molecules as a whole do not have electric charge, the component ions of the atoms themselves do, and these are subject to electrical forces. In other words, the molecules of the dielectric may be thought of as being composed of molecules

which are dipoles, fixed in position. In an electric field, the ions constituting the dielectric tend to become slightly displaced from their normal position.

If a dielectric composed of dipole molecules having random orientation is subjected to an electric field, there will be a transient or temporary rotation of the molecule dipoles so that they may align themselves with the direction of the applied field. The molecular rotation of the ions in the direction of the electric field lasts for a very short interval of time but produces a displacement current within the dielectric in the process. When the electric field is removed, the molecules are able to resume their random orientation; in so doing they produce a transient or temporary displacement current within the dielectric in a direction opposite to that which occurred when the electric field was first applied; they thus give up the energy imparted to them when the field was first applied.

87. Magnetic Behavior. Although it is possible to distinguish between at least five different kinds of magnetic materials—according to their atomic or molecular structure—ferromagnetic materials constitute the materials of greatest engineering importance; only ferromagnetic materials will be dealt with here.

In the crystal lattice of ferromagnetic materials an electron is shared by several ions. As a result, strong forces of interaction exist, and these tend to align the spins of electrons of all ions in a certain microscopic crystal domain.

Ferromagnetic materials are viewed as consisting of many small regions, called *domains*, which may be larger or smaller than a crystal grain. Within each separate domain, the atoms are presumed to be aligned in a given direction as a result of the interaction forces mentioned. In the unmagnetized state, the atoms of the different domains are randomly oriented and no net or resultant magnetic effect is observed. When a magnetic field is applied to such a ferromagnetic material, however, the magnetization increases. For small magnetizing forces, those domains in which the atoms are nearly aligned with the direction of the magnetic field are oriented in its direction. As the magnetizing force increases, other domains are aligned with it, but at a decreasing rate, until finally saturation is reached when all domains are aligned in the direction of the applied field.

References

Electromagnetism and the Structure of Matter:

Andrade, E. N. daC.: "The Structure of the Atom," Harcourt, Brace, 1927.
Bitter, F.: "Nuclear Physics," Addison-Wesley, 1950.
Born, M.: "Atomic Physics," 3d ed., Blackie, 1944.
Brown, T. B.: "Foundations of Modern Physics," Wiley, 1948.
Darrow, K. K.: "Introduction to Contemporary Physics," 2d ed., Van Nostrand, 1939.
Evans, R. D.: "The Atomic Nucleus," McGraw-Hill, 1955.
Finkelnburg, W.: "Atomic Physics," McGraw-Hill, 1950.
Gaynor, F.: "Pocket Encyclopedia of Atomic Energy," Philosophical Library, 1950.
Houston, W. V.: "Principles of Mathematical Physics," 2d ed., McGraw-Hill, 1948.
Hull, G. F.: "Elementary Modern Physics," 2d ed., Macmillan, 1949.
Oldenberg, O.: "Introduction to Atomic Physics," 2d ed., McGraw-Hill, 1954.
Rice, F. O., and E. Teller: "The Structure of Matter," Wiley, 1949.
Richtmyer, F. K., E. H. Kennard, and T. Lauritsen: "Introduction to Modern Physics," 5th ed., McGraw-Hill, 1955.
Ridenour, L. N.: "Modern Physics for the Engineer," McGraw-Hill, 1954.
Semat, H.: "Introduction to Atomic Physics," 3d ed., Rinehart, 1952.
Slater, J. C.: "Modern Physics," McGraw-Hill, 1955.
Tolansky, S.: "Introduction to Atomic Physics," Longmans, 1949.
Van Name, F. W., Jr.: "Modern Physics," Prentice-Hall, 1952.

ELECTROSTATICS

88. Definition. Electrostatics is that branch of science which deals with the laws of electricity at rest. Historically, it is the oldest branch of electricity, but it has little direct application to electrical communication, which requires that some electrical quantity vary with time if intelligence is to be transmitted. The ideas originally developed in the study of electrostatics are but special cases of more general relations of electromagnetism.

89. Electric Charges. For all engineering purposes, electrified particles are completely specified by the magnitude and sign of their electric charge and by their rest mass.

An electron is the natural elementary quantity of negative electricity or the natural elementary negative electric charge. The charge (or quantity of electricity) of an electron, represented by the symbol e, is -1.602×10^{-19} coulomb, and the rest mass of an electron is 9.107×10^{-31} kg. Alternatively, one may define the coulomb, the unit of charge for engineering purposes, as the charge whose magnitude is equal to that of 6.242×10^{18} electrons, and the kilogram as the mass represented by 1.098×10^{30} electron rest masses.

A proton is the natural elementary quantity of positive electricity or the natural elementary positive electric charge. The charge of a proton is numerically equal to the charge of an electron but is of opposite sign, or $+1.602 \times 10^{-19}$ coulomb. The rest mass of the proton is 1,835 times that of the electron, or 1.671×10^{-27} kg.

The quantity of electricity, electric charge, on (or in) a body is the excess of one kind of charge over that of the other. A plus sign indicates an excess of protons or positive ions or, conversely, a deficiency of electrons or other negative charges. Likewise, a minus sign indicates an excess of negative charges or a deficiency of positive charges.

90. Electric Field. Associated with each electric charge is a sphere of influence or a field of force, which tends to accelerate or impart motion to other charges in the field

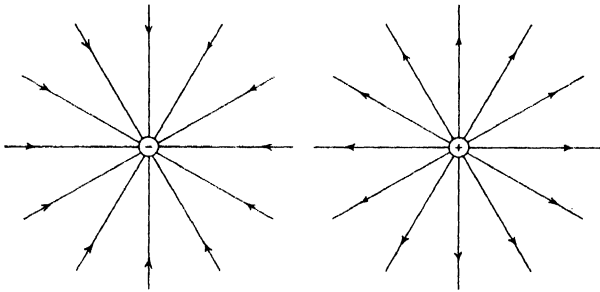



FIG. 23. Representation in a two-dimensional plane of lines of force about isolated electric charges.

of force. Such a field of force is detected only by the force it exerts on charged particles. The charge and its own field of force cannot be separated; indeed, the charge is sometimes looked upon as the point at which the field may be considered to be concentrated.

Since any charge is the result of the excess or deficiency of electrons, the charge, Q, of any substance is given by $Q = \pm ne$, where n is any integer expressing the number of elementary charges involved and e is the numerical value of the electronic charge. The sign of the charge, Q, is positive (+) or negative (−) depending upon whether the charge produces a force on other charges like that of a proton or an electron, respectively.

The sphere of influence, or the field of force, surrounding a charged particle is represented by lines of force that originate at the center of the charge and radiate outward in straight lines in all directions, as shown in Fig. 23. The direction of the lines of force designate the direction of the force that the charge exerts on an arbitrarily selected positive test charge, and the density of the lines of force is proportional to the magnitude of the force.

Since electrical effects in free space are propagated with the velocity of light, the field of any element of charge may be supposed to be emitted continuously from that element and to spread out in all directions from it with the speed of light. Such a view should not be regarded as a "true explanation" of actual physical behavior, of

course, but rather as a convenient means of accounting for the fact that electrical effects, due to the motion of charges, at a distance from the charges are observed to take place after a time interval whose magnitude depends upon the distance between the charges and the points of observation. Electrical effects at a distance are retarded in time.

91. Lines of Force and Tubes of Force. The properties of charged particles may be represented by means of lines of force, a concept to which Faraday attached physical significance. A line of force may be considered as a path indicating the direction of the force applied upon a positive test body (charge) in an electric field. A line of force in an electric field is a curve so drawn that, everywhere, it indicates the direction of the electric intensity, the sense in which the line is described being indicated by an arrowhead on the diagram to show the direction in which the positive test charge would tend to move if placed at the position of the arrowhead.

FIG. 24. Tube of force between two charged bodies.

A bundle of a large, but constant, number of lines of force is known as a "tube" of force. For a single isolated charge, the tube of force is actually a cone, but in most cases it is truly represented as a tube, as shown in Fig. 24.

92. Charge Density. A point charge, *i.e.*, an electrical charge so small that it may be considered as concentrated at a point in space, is the basis of electrostatics, and it

FIG. 25. Surface and volume charge density for electric charges distributed over a surface (left) or throughout a volume (right). The surface and volume are infinitesimal in size.

is conceivable that electrical problems could be solved in terms of point charges only. When many charges are to be dealt with, however, such a procedure becomes tremendously tedious and cumbersome and, in such cases, the concept of charge density is particularly useful in simplifying the solution of electrical problems. Two kinds of charge density of practical importance may be distinguished.

If the point charges are distributed over a surface area, as shown in Fig. 25*a*, the surface charge density, σ, is defined as the limiting ratio of the charge to the surface area, as the area approaches zero. If, in mks units,[1] Q is the charge in coulombs and

[1] In this section, electrical quantities are expressed in meter-kilogram-second (mks) units since this system was recommended for adoption after January, 1940, by the International Electrotechnical Commission. Conversion from the mks to the cgs units may be made by means of the conversion

s is the surface area in square meters, the surface charge density in coulombs per square meter is defined to be

$$\sigma = \lim_{\Delta s \to 0} \frac{\Delta Q}{\Delta s}$$

If the charges are distributed throughout a volume, V, as shown in Fig. 25b, the volume charge density, ρ, is defined as the limiting ratio of the charge to the volume, as the volume approaches zero. If Q is the charge in coulombs and V is the volume in cubic meters, the volume charge density is defined mathematically as

$$\rho = \lim_{\Delta V \to 0} \frac{\Delta Q}{\Delta V}$$

The charge density is defined at a point on a surface or in a volume, and its magnitude may vary from point to point. The sign of the charge density is that of the charge.

93. Electric Field Intensity. The electric field intensity (also called "electric intensity," "electric field strength," or "electric force") at a point is a vector whose direction is that of the force that would be exerted on a charged particle placed at the point and whose magnitude is equal to the ratio of the force to the charge of the electrified (charged) particle, provided the latter is sufficiently small so that it does not measurably affect or alter the electric field. A vanishingly small charge must, therefore, be employed as the test body to measure the electric field intensity. Hence, the electric field intensity may be defined mathematically as

$$\mathcal{E} = \lim_{\Delta Q \to 0} \frac{\Delta f}{\Delta Q}$$

where, in mks units, f is the force in newtons, Q is the charge in coulombs, and $\mathcal{E}$ is the electric field intensity in volts per meter.

94. Coulomb's Law. As early as 1773, Sir Charles Cavendish made the first quantitative measurements in the field of electricity and deduced the relation between the forces exerted by charges, but his results were unpublished and remained unknown for almost a century. In the meantime, in 1785, Charles Augustin Coulomb deduced the same relations between the forces on charges by another approach, and this important relation is now known as Coulomb's law.

Coulomb's law may be stated in words as follows: In a homogeneous, isotropic medium the force of attraction or repulsion between two charges of electricity concentrated at two points that are small compared to the distance between them is proportional to the product of their charge magnitudes, inversely proportional to the square of the distance between the charges, and inversely proportional to a characteristic constant of the medium in which the charges reside, called its "permittivity." The force between unlike charges is one of attraction; the force between like charges is one of repulsion.

Mathematically, Coulomb's law may be stated in the form

$$f = K_e K' \frac{Q_1 Q_2}{\epsilon l^2} = K_e K' \frac{Q_1 Q_2}{\epsilon_0 \epsilon_r l^2}$$

where, in mks units, f is the force in newtons exerted by the two charges on one another, Q_1 and Q_2 are the two charges in coulombs, $\epsilon = \epsilon_0 \epsilon_r$ is a physical property of the homogeneous isotropic medium known as the *permittivity* and is measured in farads per

factors given in Table 13. The question of rationalization was left open by the IEC so that, at the present time, there is no unanimity as to which set of equations should be free from the unavoidable constant, 4π, representing the number of solid radians in a closed surface. In this section, K coefficients, with suitable primes, are used to replace coefficients involving the factor 4π. By inserting the appropriate value of K, as given in Table 6, the equations in this section are directly applicable to either the unrationalized or the rationalized system.

meter, ϵ_0 is the permittivity of free space (vacuum) measured in farads per meter, ϵ_r is a numeric expressing the ratio of the permittivity of any substance to the permittivity of free space, l is the distance between the point charges in meters, K_e is a constant which depends upon the units in which the force and charge are measured, and K', whose value is given in Table 6, is a constant whose value depends upon whether the unrationalized or the rationalized system of equations is employed.

Table 6. Factors for Rationalized and Unrationalized Systems of Units

Factor	Value in unrationalized system	Value in rationalized system
K'	1	$\frac{1}{4\pi}$
K''	$\frac{1}{4\pi}$	1
K'''	4π	1
K''''	1	4π

Coulomb's law provides a basis for relating mechanical units (of force) with electrical units (of charge). Another important result of Coulomb's law and the specification of the charge in terms of multiples of the elementary charge ($Q = ne$) is that the mechanical forces produced by charges may be accounted for by the linear superposition of charges, for multiplying one of the charges by any quantity (keeping the other charge constant) multiplies the resulting force by the same quantity. It makes no difference which charge is kept constant and which is varied; hence, it follows that charges may be combined in any order. This statement applies whether the charges are stationary or in motion so that the principle of linear superposition also applies to electric currents (which are charges in motion) so long as the medium is homogeneous and isotropic.

95. Conservative Properties of Static Fields. Problems in electrostatics may be solved by the fundamental relations expressed by Coulomb's law, which deals with forces that have direction as well as magnitude. The calculation of directed forces is a laborious and tedious undertaking, so that, except in the simplest configurations of charges, calculations by Coulomb's law become too cumbersome for general use. The concept of potential greatly simplifies problems in electricity by eliminating the need to consider direction (except in a positive and negative sense). Since the potential concept is applicable only to conservative fields, it is necessary to ascertain that the electrostatic field is conservative, *i.e.*, that the energy of an electrostatic field is a function of position only and does not depend upon the path that the charge takes between the two points.

According to Coulomb's law and the definition for work, the increment of work done in moving a charge a distance dr is given by the equation[1]

$$dw = -\mathbf{f}\cos\theta\, dl = -\mathbf{f}\cdot dr = -(K_e K' Q_1 Q_2/\epsilon r^2)\, dr$$

[1] When vector quantities are encountered, it is frequently necessary to express physical quantities in terms of the component of the vector that is along a specified path or is normal to a specified surface. For example, the increment of work is equal to the product of the distance through which a body moves by the component of force in the direction of motion, or $dw = \mathbf{f}\cos\theta\, dl$, where $\mathbf{f}$ is the force, dl is the element of length in the direction of motion, and θ is the angle between the direction of the force and the direction of motion. Thus, $\mathbf{f}\cos\theta$ is that component of the force in the direction of motion along dl. In cases such as this, the desired vector (in this example, the element of work) may be expressed in any of the alternative ways

$$\mathbf{f}\cos\theta\, dl = \mathbf{f}\cos(\mathbf{f}, l)\, dl = \mathbf{f}\cdot dl$$

where the notation $(\mathbf{f}, l)$ represents the angle between the direction of $\mathbf{f}$ and that of l. The dot (·) is used to represent the cosine of this angle, or $\cos\theta$.

The vector $\mathbf{f}$ may be resolved into its components along any set of three coordinates in space. Each such component is, of course, a vector quantity and may be specified in terms of its component vectors in the direction of each coordinate. Thus, the vector $\mathbf{f}$ may be resolved, for example, into its rectangular coordinate components, $\mathbf{f}_x$, $\mathbf{f}_y$, and $\mathbf{f}_z$, as in the expression

$$\mathbf{f} = \mathbf{f}_x + \mathbf{f}_y + \mathbf{f}_z$$

A vector may also be specified in terms of a unit vector in the direction of each coordinate and a scalar quantity representing the magnitude of the vector component along each coordinate. Thus, if $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors in the x, y, and z directions, respectively, and if f_x, f_y, and f_z are the magnitudes

the negative sign being used because positive work is done in moving the charge in the $-r$ direction. The total amount of work done in moving the charge from any point A to any other point B whose distances from any reference point are r_A and r_B, respectively, is then the integral of dw between the two points, or

$$w = w_A - w_B = \int_B^A dw = -\int_{r_B}^{r_A} \left(\frac{K_e K' Q_1 Q_2}{\epsilon r^2}\right) dr = -\left(\frac{K_e K' Q_1 Q_2}{\epsilon r}\right)\left(\frac{}{r_A} - \frac{1}{r_B}\right)$$

This result is independent of the path taken between A and B and depends only upon the positions of the end points of the path traversed. Therefore, the electrostatic field is conservative.

96. Electrostatic Potential. The electric potential difference between two points in an electrostatic field is equal to the work associated with the transfer of unit quantity of positive charge from one point to another. The maximum amount of work is done by (or on) charges that move along the lines of electric force, whereas no work is done by (or on) a charge that moves perpendicular (normal) to lines of electric force. If A and B are the two points between which a unit positive charge is transferred, then B is at a higher potential than A if external energy is required to transfer unit positive charge from A to B.

The great advantage of the concept of potential is that, through its use, vector calculations of the forces acting on charges may be resolved into scalar calculations which are much simpler to manipulate. From the definition given above, it is evident that there exists only a means for determining the difference or change of potential between two points and not for determining the absolute value of the potential of any point. If it is possible to establish as an arbitrary reference an equipotential surface, *i.e.*, a surface normal to the lines of force upon which no work is required to transfer a positive charge from one point to another on the surface, this surface may be established or defined as being at zero potential. The potential of any point may then be specified relative to the equipotential surface arbitrarily designated as being at zero potential. In engineering work the earth is often arbitrarily taken as the zero reference equipotential surface, and potential differences are then measured relative to the potential of the earth's surface. In other cases, particularly in theoretical work, it is convenient to assign zero potential to a sphere of infinite radius, since the force produced at a point infinitely removed from a charge is zero.

The increment of work per unit charge that is done in moving the charge a distance dl in an electric field is $dw_Q = -(f/Q)\cos\theta\, dl$. But by definition, $\mathcal{E} = f/Q$ and, therefore, $dw_Q = -\mathcal{E}\cos\theta\, dl$. The difference in potential between two points A and B is the integral along the path from A to B, or

$$E_{AB} = V_A - V_B = -\int_B^A \mathcal{E}\cos\theta\, dl = -\int_B^A \mathcal{E}\cdot dl$$

(scalar quantities) of **f** in these three directions, the vector may be expressed in the alternative form

$$\mathbf{f} = \mathbf{i}f_x + \mathbf{j}f_y + \mathbf{k}f_z$$

The vector may represent a quantity flowing through a surface. For example, the amount of liquid that passes outward through the element of surface, ds, in unit time at any point on the surface is the product of the outward normal (or perpendicular) component of the vector **F** and the surface element, ds. If **n** is a unit vector normal to the surface elements, ds, the outward flow of liquid may be expressed in any of the alternative ways

$$\mathbf{F}\cos\theta\, ds = \mathbf{F}\cos(\mathbf{F}, n)\, ds = \mathbf{F}\cdot n\, ds = \mathbf{F}_n\, ds$$

where $(\mathbf{F}, n)$ represents the angle between the direction of **F** and the surface normal, n. The dot (·) expresses the cosine of the angle between **F** and n, and $\mathbf{F}_n$ is the component of **F** perpendicular to the surface element.

In some texts, especially those stressing vector notation, vector quantities are printed in boldface (heavy) type, as has been done for illustration in this footnote. In other cases, as in the main text of this section, ordinary light-face characters are employed. In this section script letters have been used for those electrical quantities (as $\mathcal{E}$, $\mathfrak{D}$, $\mathcal{H}$, and $\mathcal{B}$) which are regarded as vectors in space.

where V_A is the absolute potential of point A, V_B is the absolute value of potential of point B, $E_{AB} = V_A - V_B$ is the difference in potential between points A and B, $\mathcal{E}$ is the electric intensity in the region along the path from A to B, and θ is the angle between the direction of $\mathcal{E}$ and that of dl. Although absolute values for V_A and V_B are not known, the difference $V_A - V_B$ can be ascertained.

If A and B are the two points between which a positive charge is transferred, then B is at a higher potential than A if external work is done in transferring the charge from A to B, *i.e.*, if energy is expended to effect this transfer. Point B is then said to be at a higher potential than point A; alternatively, point B is said to be positive with respect to point A. Energy is given up when the positive charge is transferred from B to A.

In the mks system of units, unit potential difference is the joule per coulomb, which has been given the name *volt*.

If the electrostatic field is made up of concentrated charges $Q_1, Q_2, Q_3, \ldots, Q_n$ in a homogeneous isotropic medium whose permittivity is ϵ and in which the distances of these charges from a point P are $r_1, r_2, r_3, \ldots, r_n$ and if the potential of a sphere infinitely distant is regarded as the reference or zero potential, then, with respect to the zero reference potential, the potential of the point P is

$$E_P = -\int_\infty^{r_1} \frac{K_eK'Q_1}{\epsilon r_1^2}\,dr - \int_\infty^{r_2} \frac{K_eK'Q_2}{\epsilon r_2^2}\,dr - \int_\infty^{r_3} \frac{K_eK'Q_3}{\epsilon r_3^2}\,dr - \cdots - \int_\infty^{r_n} \frac{K_eK'Q_n}{\epsilon r_n^2}\,dr$$

$$= -\frac{K_eK'}{\epsilon}\left(\frac{Q_1}{r_1} + \frac{Q_2}{r_2} + \frac{Q_3}{r_3} + \cdots + \frac{Q_n}{r_n}\right)$$

$$= -\frac{K_eK'}{\epsilon}\sum_{n=1}^{m} \frac{Q_n}{r_n}$$

In practice such an expression as that given above is applicable only if a very small number of charges is encountered, for the calculations become tedious and cumbersome if many charges must be taken into account.

When many charges make up an electrostatic field, it is often convenient to disregard the individual charges and deal with their net or over-all effects as expressed by the density of charge. If σ is the surface charge density and ρ is the volume charge density, then the charge distributed over an elementary area, ds, is $dQ = \sigma_n\,ds$ and the charge distributed throughout an elementary volume, dv, is $dQ = \rho\,dv$. In the case of a continuous distribution of charge throughout a volume and also over a surface, the potential of the point P (with respect to zero potential at a point infinitely remote), whose distance from the volume and the surface is r, is given by

$$E_P = \iint_s \frac{dQ}{r} + \iiint_v \frac{dQ}{r} = \iint_s \frac{\sigma_n\,ds}{r} + \iiint_v \frac{\rho\,dv}{r}$$

The electrostatic potential (defined as the work per unit charge done in, or available for, transporting electric charges) is applicable to static fields. The concept of potential may be generalized and applied to fields in which the charges vary with time. The potential field is a scalar field whereas the field of force is a vector field.

97. Potential Gradient. The gradient of a scalar field at a point is a vector whose magnitude is equal to the space rate of change of the scalar field in the direction of greatest increase and whose direction is that of the greatest increase. This general definition may be applied to the potential field as follows:

From the relations given in Sec. 96 it is easy to show that $dE = -\mathcal{E}\cos\theta\,dl$, from which $\mathcal{E}\cos\theta = -(dE/dl)$. If the direction of the field of force exerted on a small positive test charge is in the direction of the element of path length, dl, then $\cos\theta = 1$. For this condition the change of potential is a maximum with respect to a change of position. The maximum rate of change of potential with position is a vector quantity

whose direction is at right angles to the equipotential surface or along the electric lines of force; it is called the "potential gradient" and is abbreviated "grad E." Hence, from the results above,

$$\text{grad } E = \frac{dE}{dl} = -\mathcal{E}$$

In the mks system of units, the gradient is measured in volts per meter.

The potential gradient may be resolved into orthogonal components. In rectangular coordinates,

$$\text{grad } E = \nabla E = \left(\mathbf{i}\frac{dE_x}{dx} + \mathbf{j}\frac{dE_y}{dy} + \mathbf{k}\frac{dE_z}{dz}\right) = -(\mathbf{i}\mathcal{E}_x + \mathbf{j}\mathcal{E}_y + \mathbf{k}\mathcal{E}_z)$$

where **i**, **j**, and **k** are unit vectors in the x, y, and z directions, respectively, and the components of E and $\mathcal{E}$ in each of the rectangular coordinate directions are indicated by the subscript x, y, or z.

A pictorial representation of potential gradient resolved into its orthogonal components in rectangular coordinates is given in Fig. 26. The potential gradient is, strictly speaking, applicable to the field intensity only in the case of an electrostatic field.

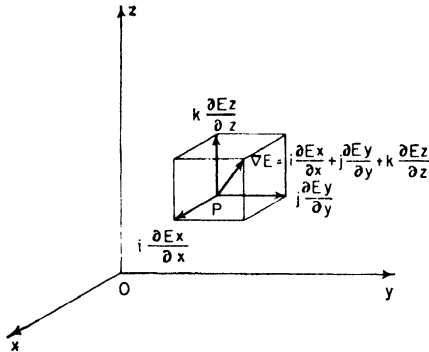



FIG. 26. Potential gradient at point P in terms of its orthogonal components, expressed in rectangular components.

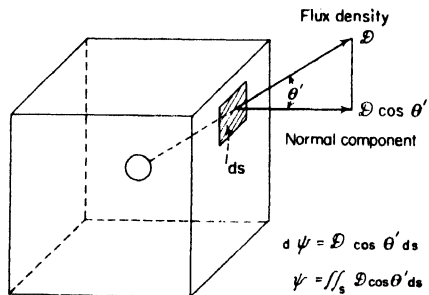



FIG. 27. Determination of electric flux diverging from a closed surface. The total flux is the sum of all components normal to a closed surface.

98. Electric Flux. Consider a surface element at a point in an electric field for which the electric intensity is $\mathcal{E}$. If θ represents the angle between the direction of $\mathcal{E}$ and the normal to the surface, **n**, then, as shown in Fig. 27 the component of $\mathcal{E}$ perpendicular to the surface is $\mathcal{E} \cos \theta = \mathcal{E} \cos (\mathcal{E}, n) = \mathcal{E} \cdot n = \mathcal{E}_n$. The product of the normal component of the electric field intensity and the surface is called the "electric flux." The electric flux, represented by the symbol ψ, may be represented by lines that are drawn in the same direction as the lines of electric intensity but of such spacing that their density (in lines per unit area) perpendicular to the direction in which the lines run is equal to the dielectric flux density at the point at which the flux is measured. From a physical point of view, the electric flux may be regarded as the number of lines of electric field that cross perpendicular to a surface; it is the total number of such lines and not the density in lines per unit area.

Electric flux is a scalar quantity, the sign of which depends upon the assumption made concerning the positive direction of the normal. With a closed surface, the outward normal is taken as positive. For an open surface, if the positive sense is that in which the periphery of the surface is described by the rotation of a right-hand screw normal to the surface, the positive direction of the normal is that in which the screw advances.

In the mks system of units, the electric flux, ψ, is measured in coulombs, the same unit in which charge is measured. In the unrationalized system, the magnitudes of the units of electric flux and charge are the same, but in the rationalized system the unit of electric flux is $1/4\pi$ as large as in the unrationalized system.

99. Electric Displacement; Electric Flux Density. The electric displacement, or the electric flux density, at any point in an isotropic dielectric medium is a vector having the same direction as that of the electric field intensity, and a magnitude that is ϵ times that of the electric field intensity. If, in mks units, $\mathcal{E}$ is the electric field intensity in volts per meter, $\mathfrak{D}$ is the electric displacement in coulombs per square meter, and ϵ is the permittivity of the dielectric medium in farads per meter, then the electric flux density in an isotropic dielectric is

$$\mathfrak{D} = K''\epsilon\mathcal{E}$$

Thus, the displacement takes into account the permittivity properties of the medium through which the lines of electric flux flow; the properties of the medium are inherent in the displacement vector itself.

If $\mathfrak{D}$ represents the electric displacement vector making an angle θ with the surface through which the electric field passes, then the element of electric flux crossing an element of surface is

$$d\psi = \mathfrak{D} \cos \theta \, ds = \mathfrak{D} \cos (\mathfrak{D}, s) \, ds = \mathfrak{D}_n \, ds = \mathfrak{D} \cdot ds$$

The total flux is the integral over the closed surface through which flux passes, or

$$\psi = \iint_s \mathfrak{D} \cos \theta \, ds = K'''Q$$

where, in mks units, ψ is the electric flux in coulombs, $\mathfrak{D}$ is the displacement in coulombs per square meter, and ds is an element of surface, expressed in square meters. This equation merely states that the total flux diverging from a closed surface is proportional to the total charge enclosed by the surface, as shown in Fig. 28.

100. Dielectric Polarization. For a material dielectric the displacement, $\mathfrak{D}$, may be split into two parts

$$\mathfrak{D} = \epsilon\mathcal{E} = \epsilon_0\epsilon_r\mathcal{E} = \epsilon_0\mathcal{E} + (\epsilon_r - 1)\epsilon_0\mathcal{E}$$

The first term, $\epsilon_0\mathcal{E}$, is the displacement for free space. The second term, $(\epsilon_r - 1)\epsilon_0\mathcal{E}$, represents an actual displacement or shift of bound electric charges within the material dielectric itself. This latter displacement or shift is customarily measured in units of charge per unit area, or in mks units, in coulombs per square meter; it is called the "polarization" of the dielectric and is designated by the symbol $\mathcal{P}$. Since the unit of displacement is K' times the unit of charge, $(\epsilon_r - 1)\epsilon_0 = 4\pi$ so that

$$\mathfrak{D} = \epsilon_0\mathcal{E} + 4\pi\mathcal{P}$$

101. Gauss's Theorem. Coulomb's law is one of the most important of all electrical laws and one way of expressing the relation between charge and force has already been given for an isotropic, homogeneous medium. The inverse-square law, already given, is the form that is most convenient in providing a quantitative idea of the forces between isolated electric charges, but Coulomb's law may also be given in alternative forms that bring out certain additional general and important properties of charged bodies and facilitate many kinds of calculations. Gauss's theorem is one highly important alternative expression for Coulomb's law. The diagram of Fig. 28 illustrates the theorem.

Gauss's theorem makes use of the concept of electric flux. The theorem may be stated in words as follows: The integral, over any closed surface, of the normal component of the electric flux density or displacement is equal to the charge contained

within the surface. In mathematical notation this may be expressed as

$$\psi = \oint\!\!\oint_s \epsilon \mathcal{E} \cos \theta \, ds = \oint\!\!\oint_s \mathfrak{D} \cos \theta \, ds = \iiint_v \rho \, dv = K'''Q$$

where $\mathfrak{D}$ is the electric flux density in coulombs per square meter, $\mathcal{E}$ is the electric field intensity in volts per meter, θ is the angle between the direction of $\mathfrak{D}$ or $\mathcal{E}$ and the normal to the surface, ds is an element of surface expressed in square meters, ρ is the volume charge density in coulombs per cubic meter, and dv is an element of volume expressed in cubic meters.

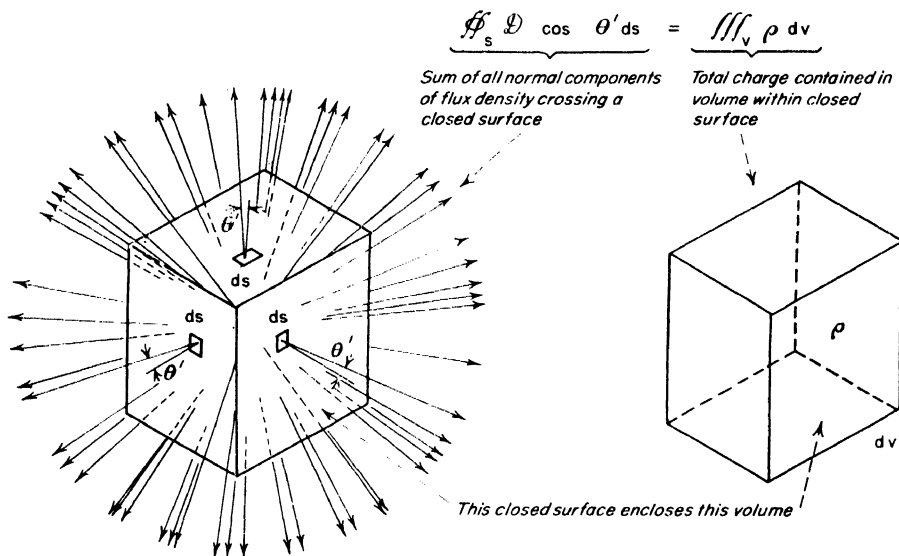



FIG. 28. Total electric flux diverging from a closed surface (left) is equal to the total charge enclosed by that surface.

The above equation is the integral form of Gauss's theorem, which may be derived as follows. If the electric displacement, $\mathfrak{D}$, crosses a surface element, ds, the normal component of flux density crossing the surface is

$$d\psi = \mathfrak{D} \cos \theta \, ds = \epsilon \mathcal{E} \cos \theta \, ds = [Q \, (\cos \theta)/r^2] \, ds$$

But $[(\cos \theta)/r^2] \, ds$ is an elementary solid angle, $d\Omega$, and, hence, $d\psi = Q \, d\Omega$. Hence the total flux is the integral over a closed surface, and since there are 4π radians in a closed surface, the flux is given by

$$\psi = \oint\!\!\oint_s Q \, d\Omega = K''Q$$

102. Equations of Laplace and Poisson. Gauss's theorem may be put into differential as well as integral form to obtain either the Laplace or the Poisson equation. The differential form may be obtained by resolving the electric flux into three orthogonal components, expanding each component in a Taylor's series expansion and forming the algebraic sum of the linear terms in this expansion for each orthogonal component, over a closed surface. If the volume enclosed is free from charge, the result of such an operation is the scalar quantity given by Laplace's equation,

$$\text{div } \mathfrak{D} = \frac{d\mathfrak{D}_x}{dx} + \frac{d\mathfrak{D}_y}{dy} + \frac{d\mathfrak{D}_z}{dz} = 0$$

whereas if the enclosed volume has a volume charge density, ρ, the equation of Poisson results,

$$\text{div } \mathfrak{D} = \frac{d\mathfrak{D}_x}{dx} + \frac{d\mathfrak{D}_y}{dy} + \frac{d\mathfrak{D}_z}{dz} = K'''\rho$$

The equations of Laplace and Poisson may be expressed in terms of potential as well as of electric field intensity since $\mathcal{E} = -(dE/dl)$. Hence the above equations become

$$\frac{d^2E_x}{dx^2} + \frac{d^2E_y}{dy^2} + \frac{d^2E_z}{dz^2} = 0$$

$$\frac{d^2E_x}{dx^2} + \frac{d^2E_y}{dy^2} + \frac{d^2E_z}{dz^2} = -\frac{K'''\rho}{\epsilon}$$

Like the integral forms for expressing Gauss's theorem, the equations of Laplace and Poisson are merely alternative ways of expressing the fundamental laws of charge which were first quantitatively observed by Cavendish and Coulomb.

103. Divergence. Divergence is a concept that is useful in describing the flow of a quantity. The divergence of a vector field at any point may be defined as a scalar quantity equal in magnitude to the limiting ratio of the outward flux, passing through a closed surface that surrounds the point, to the volume enclosed within the surface, as the volume approaches zero. Physically the operation denoted by divergence yields a measure of the flux produced per unit volume, at a point. When flux flows away from a point, divergence is indicated; when it flows toward a point, negative divergence (sometimes called "convergence") is indicated.

If, in a vector field, the vector **A** is expressed in terms of its three rectangular orthogonal coordinates, A_x, A_y, and A_z, given as functions of x, y, and z coordinates, respectively, the divergence of the vector field is a scalar quantity formed by the sum of the derivatives of each orthogonal component in the direction of its axis, or

$$\text{div } \mathbf{A} \equiv \nabla\cdot\mathbf{A} \equiv (\nabla\mathbf{A}) = \frac{dA_x}{dx} + \frac{dA_y}{dy} + \frac{dA_z}{dz}$$

$$= \left(\mathbf{i}\frac{d}{dx} + \mathbf{j}\frac{d}{dy} + \mathbf{k}\frac{d}{dz}\right)(\mathbf{i}A_x + \mathbf{j}A_y + \mathbf{k}A_z)$$

As an example of this concept, the divergence of the electric intensity, which is proportional to the volume density of charge at the point, may be cited.

104. Curl. Curl is a concept useful in describing the rotation or swirl of a flowing quantity. The curl (also called "rotation," especially by European authors) of a vector field at a point is a vector whose magnitude is equal to the limiting ratio of the circuitation around a surface element on which the point is located, to the area of the surface as the latter approaches zero, provided the surface is so oriented as to give the maximum values of circuitation. The positive direction of this vector is that traveled by a right-hand screw turning about an axis normal to the surface element when the turning of the screw gives a positive value to the circuitation or rotation.

If the vector **A** of a vector field is expressed in terms of its three rectangular components, A_x, A_y, and A_z, given as functions of x, y, and z, respectively, the curl of the vector field is the vector sum of the partial derivatives of each component with respect to the axes perpendicular to it. Mathematically,

$$\text{curl } \mathbf{A} \equiv \nabla \times \mathbf{A}$$

$$= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} = \mathbf{i}\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) + \mathbf{j}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) + \mathbf{k}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)$$

where **i**, **j**, and **k** are unit vectors along the x, y, and z axes, respectively.

As an example of the curl concept, the magnetic intensity at a point within an electric conductor, which is proportional to the current density, may be cited.

MAGNETOSTATICS

105. Definition. Magnetostatics is the branch of science that deals with the laws of permanent magnets. Prior to the discovery by Oersted in 1820 of the relation between electricity and magnetism, the study of magnetostatics was pursued independent of that of electrostatics, for in the static case no relation between electrostatics and magnetostatics can be made at all apparent. With the discovery that the flow of electric currents produced magnetic fields, there arose an attempt to explain all magnetic phenomena in terms of the flow of an electric current. Some authors have justified this point of view on the ground that, although the elementary electric charge has been discovered, no corresponding magnetic charge or pole has been observed as an entity of nature.

There is ample evidence that magnetic effects are the result of the flow of an electric current and possibly some justification for omitting a treatment of magnetostatics. Such a course overlooks the fact that many existing texts already employ the concepts of magnetostatics; it also neglects the historical approach which has given rise to much confusion in the establishment of electrical units.

106. Magnets. A magnet is a body or substance that produces a magnetic field external to itself. The first known magnets were natural magnets of lodestone, but today magnets for practical purposes are made of iron, steel, or alloys in which the principal constituents are usually iron, nickel, cobalt, and perhaps small proportions of other metals such as aluminum or copper.

A magnetic field is a vector field of force imparting, or tending to impart, motion to fine particles of a magnetic substance (*e.g.*, iron filings) situated in the magnetic field.

107. Magnetic Pole. The magnetic poles (or simply the poles) of a magnet are those portions of a magnet toward which the external magnetic field tends to converge or from which it tends to diverge. Magnetic poles always occur in pairs; they cannot be isolated as can electric charges. The size and position of the poles of a magnet depend upon its shape and also upon its magnetic characteristics. For a magnet that is long and slender (as a steel needle), the poles are located at small regions near the ends. In such a case the poles are virtually points, and it is convenient to regard them as concentrated at points.

A study of magnetostatics may be built up on the hypothesis that magnetic poles can be considered to exist in complementary pairs. The mathematics of magnetostatics then becomes very similar to the mathematics of electrostatics, but a different physical interpretation of the equations is required because isolated magnetic poles have not been identified.

108. Coulomb's Law for Magnetic Poles. Experiment shows that for long, thin, magnetized needles, the poles may be regarded as being concentrated at points near their ends. The regions at which the magnetic effects are concentrated may then be regarded as isolated magnetic poles each of which establishes its own spherical magnetic field of force, much as isolated charges establish their spherical electrical field of force. This point of view is justified by the mathematical convenience of the equations of magnetostatics (and the similarity of the equations of magnetostatics to those of electrostatics) rather than by the precision with which this view accords with physical facts. If such a view is adopted, the force between magnetic poles obeys the inverse-square law, and Coulomb's law may be applied to magnetic poles as well as to electric charges.

Coulomb's law for magnetic poles may be stated as follows: In a homogeneous, isotropic medium, the force between two magnetic poles that are small compared to the distance between them is proportional to the product of the magnitudes (or strengths) of the two poles, inversely proportional to the square of the distance between them, and inversely proportional to a characteristic property of the medium in which the

poles reside, called its "permeability." The force between unlike poles is one of attraction; the force between like poles is one of repulsion.

Mathematically, Coulomb's law for magnetic poles may be stated as

$$f = K_m K' \frac{M_1 M_2}{\mu l^2} = K_m K' \frac{M_1 M_2}{\mu_0 \mu_r l^2}$$

where, in mks units, f is the force in newtons exerted between the two poles, M_1 and M_2 are the magnitudes of the pole strengths of the two magnetic poles, $\mu = \mu_0 \mu_r$ is a physical property of the homogeneous, isotropic medium in which the poles reside, called the *permeability* of the medium, μ_0 is the permeability of free space (vacuum) measured in henrys per meter, μ_r is a numeric expressing the ratio of the permeability of any material substance to that of free space, l is the distance between the two poles in meters, K_m is a constant relating pole strength and mechanical units of force, and K' is a constant which adapts the equation to the rationalized or unrationalized system and whose value is given in Table 6.

With the exception that no free, isolated, magnetic poles exist (so that, for example, there is no magnetostatic equivalent of the electrostatic charge density), the reasoning developed for electric charges applies to magnetic poles. Thus, Gauss's theorem and the equation of Laplace apply to magnetic poles as well as to electric charges, but Poisson's equation becomes identical with Laplace's since there is no magnetic "pole density."

109. Pole Strength. The strength of a magnetic pole that can be considered as concentrated at a point is measured by the force exerted on the pole when it is placed in a magnetic field of known intensity in a vacuum. A unit magnetic pole is one which is concentrated at a point and which has such strength that, when placed at a unit distance from an exactly similar pole, it will experience a unit repelling force. In the cgs electromagnetic units, the unit of pole strength is the pole, but this unit has no name in the cgs electrostatic or in the mks system of units.

110. Magnetic Field Intensity. The magnetostatic field intensity due to a magnetic pole is defined as a vector whose magnitude is equal to the force per unit magnetic pole strength exerted on a test body, and whose direction is the same as the direction of the force exerted on the test body. The test body is a concentrated north-seeking pole of such small strength (strictly, infinitesimal strength) that its presence does not alter the magnetic field that it is intended to measure.

The magnetic field intensity may be defined by the magnetostatic relation,

$$\mathcal{H} = \lim_{\Delta M \to 0} \frac{\Delta f}{\Delta M}$$

where, in the mks system of units, $\mathcal{H}$ is measured in ampere-turns per meter, f is the force in newtons, and M is the pole strength. This equation is analogous to the electric field intensity produced by static electric charges. The magnetic field intensity may also be defined by a dynamic relation in terms of current flow (see Sec. 127).

The direction and magnitude of the magnetic field intensity may be represented by lines of magnetic force, just as lines of electric force represent the direction and magnitude of the electric field intensity.

111. Magnetic Flux. The magnetic flux, ϕ, produced by magnetic poles, through a surface may be regarded as the number of lines of magnetic force that cross perpendicular to the surface. It is a scalar quantity, derived in a way completely similar to that by which electric flux is derived. In the mks system of units, the magnetic flux is measured in webers.

112. Magnetic Induction; Magnetic Flux Density. The magnetic induction, or the magnetic flux density, at any point in a magnetic field is a vector whose magnitude is μ times that of the magnetic field intensity, $\mathcal{H}$, and whose direction is that of the magnetic field intensity. If $\mathcal{H}$ is the magnetic field intensity and μ is the permeability

of the magnetic medium, then the magnetic induction (or magnetic flux density) is given by the relation

$$\mathcal{B} = \mu\mathcal{H}$$

In mks units the magnetic induction is measured in webers per square meter.

In terms of the magnetic flux density, the magnetic flux is given by

$$\phi = \iint_s \mathcal{B} \cos\theta \, ds$$

where θ is the angle between the direction of $\mathcal{B}$ and the normal to the surface element. Over a closed surface,

$$\phi = \oiint_s \mathcal{B} \cos\theta \, ds = 0$$

This relation may be regarded as a result of the absence of free, isolated magnetic poles. Permanent magnets, however small, are always found in equal and opposite pairs. Therefore, for any volume containing a permanent magnet, as many lines of magnetic force leave the surface of enclosure as enter it; so the net or resultant flux is zero. In mks units, ϕ is the magnetic flux in webers, $\mathcal{B}$ is the magnetic flux density in webers per square meter, ds is an elementary surface area expressed in square meters, and θ is the angle between the direction of the flux density and the normal to the element of surface, ds.

In differential form, Gauss's theorem for magnetic flux density is given by

$$\frac{d\mathcal{B}_x}{dx} + \frac{d\mathcal{B}_y}{dy} + \frac{d\mathcal{B}_z}{dz} = 0$$

In words, the divergence of magnetic flux from a closed surface is zero.

113. Magnetization. Magnets are regarded as being produced by the proper alignment or orientation of elementary magnets (of crystalline dimensions) within the crystalline structure of the magnetic material. In the natural state the elementary magnets are at random orientation; hence, the substance shows no magnetic properties. As more and more of the elementary magnets assume the same orientation, the magnetism increases, up to the limit attained when all elementary magnets are aligned in the same direction.

The magnetic properties of material substances may be separated from those of free space in the same way in which dielectric polarization was separated from the permittivity of free space. Thus, we may write

$$\mathcal{B} = \mu\mathcal{H} = \mu_0\mu_r\mathcal{H} = \mu_0\mathcal{H} + (\mu_r - 1)\mu_0\mathcal{H} = \mu_0\mathcal{H} + 4\pi\mathfrak{M}$$

where $\mathfrak{M}$ is the intensity of magnetization, which, in the mks system, is measured in webers per square meter.

114. Permeability and Reluctivity. Permeability, designated by the symbol μ, is the property of an isotropic medium which determines, under specified conditions, the magnitude relation between magnetic induction and magnetic intensity (also called "magnetizing force") of the medium.

Under the specified conditions, permeability is measured as the ratio of the magnetic induction to the magnetic intensity. Mathematically, the permeability is defined to be

$$\mu = \frac{\mathcal{B}}{\mathcal{H}}$$

Reluctivity of a medium is the reciprocal of its permeability. Mathematically,

$$\nu = \frac{1}{\mu} = \frac{\mathcal{H}}{\mathcal{B}}$$

ELECTRODYNAMICS

115. Definition. Thus far we have dealt with electric charges and magnetic poles in the static case; *i.e.*, we have dealt with charges and poles free from motion relative

to the reference coordinate system. So long as only static conditions were studied, there were no discernible connections between electricity and magnetism. With the discovery, in 1820, by Oersted (quickly repeated by Ampère) that the flow of an electric current produced a magnetic field, and the further discovery by Faraday in 1831 that a varying magnetic field was capable of producing an induced electromotive force, the connection between electricity and magnetism was clearly and firmly established, and from then on the unified subject of electromagnetism was studied.

Electromagnetism may be defined as the study of electricity (electric charges) in motion, the magnetic effects produced by the flow of electricity, and, conversely, the electrical effects arising from the relative motion of a magnetic field and path through which electric current may pass or flow. The current may be a conduction current or a displacement current. In the former case, electromagnetic energy may be guided between two points connected by a conductor. In the latter case, electromagnetic energy may be transferred from one point to another through a dielectric.

The term "electrodynamics" (or electrokinetics) is used in place of electromagnetism, particularly when the motional aspect of the electric charges is emphasized, as in distinguishing electrostatics from electrodynamics. The term "electromagnetism" usually includes electrostatics and, hence, is often regarded as a more encompassing term than "electrodynamics."

116. Electric Current. An electric current through a surface is the time rate at which positive or negative charges pass through it. If both positive and negative charges are simultaneously passed through the surface, the current is the time rate of passage of the algebraic sum of the two. Three types of current flow are recognized: conduction current, convection current, and displacement current.

A conduction current comprises the movement of negative charges (electrons) exclusive of any movement due to the transportation of negative electricity by masses larger than electrons. A convection current is a current in which the charges are electrified particles larger than electrons. The displacement current through any surface in an isotropic dielectric is proportional to the time rate of change of the displacement flux through the surface.

In accordance with these definitions, the conduction and convection currents may be regarded as the flow of physical particles through a surface, the magnitude of the current being proportional to the net flow of the charged particles transported through the surface. No charge can pass through an ideal dielectric, and yet the time-varying charges accumulating on the plates of a capacitor give every indication that their physical presence is equivalent to the flow of a time-varying current through the dielectric. The magnitude of this displacement current is proportional to the time rate of change of the electric flux between the two conductors of the capacitor. But such a time rate of change of electric flux exists in the case of conduction and convection currents as well. Instead of visualizing the electric current as the flow of actual particles through a surface, a more fundamental and basic point of view is to regard the electric current as the time rate of change of the electric flux produced by moving charges. If the charges are able to pass through a surface, then the effect is the usual one associated with a conduction current or with the continuous passage of charged particles. On the other hand, if the charges are prevented from crossing a surface (as in the case of the dielectric in a capacitor), the flow of displacement current is still permissible without any change in fundamental concepts. In any case, the total current flowing at any instant of time is the sum of the conduction, convection, and displacement components of current.

If Q represents the charge of moving charged particles, the conduction (or convection) current is given by

$$I_c = \frac{dQ}{dt}$$

whereas if $\psi = K'''Q = Q/K'$ is the electric flux produced by the charge Q and flowing through a surface, the displacement current through this surface due to the variation

of the electric flux is

$$I_d = K' \frac{d\psi}{dt}$$

The total current, composed of conduction (or convection) and displacement components, is

$$I_t = I_c + I_d = \frac{dQ}{dt} + K' \frac{d\psi}{dt}$$

In mks units, the charge and flux are measured in coulomb units, and the current (all components, individually or collectively) is measured in coulombs per second or in amperes.

117. Current Density. Current density at a point is a vector whose direction is the same as that of the current flow and whose magnitude is equal to the limiting ratio of the current flowing normally across the surface, as the surface becomes infinitesimally small. The current density represents the current flow per unit area at the point. In mediums that obey Ohm's law, the current density is proportional to the electric field intensity.

If s is the area surrounding the specified point at which the current flows, I is the magnitude of the current flowing, θ is the angle between the direction of current flow and the normal to the surface through which it flows, then the normal component of current across the surface is $I_n = I \cos \theta$, and the current density is defined to be

$$\mathcal{J} = \lim_{\Delta s \to 0} \frac{\Delta I_n}{\Delta s}$$

In the mks system of units, current is measured in amperes, area is measured in square meters, and current density is, therefore, measured in amperes per square meter.

In terms of the current density, the total current flowing across a surface, s, is

$$I = \iint_s \mathcal{J} \cos \theta \, ds = \iint_s \mathcal{J}_n \, ds = \iint_s \mathcal{J} \cdot ds$$

If $\mathcal{J}_c$ is the conduction (and convection) component of the current density, and $\mathcal{J}_d = K'(d\mathfrak{D}/dt)$ is the displacement component, the total current density is then $\mathcal{J}_T = \mathcal{J}_c + \mathcal{J}_d = \mathcal{J}_c + K'(d\mathfrak{D}/dt)$, and the normal component, flowing perpendicularly through a surface, is given by

$$\mathcal{J}_n = \left(\mathcal{J}_c + K' \frac{d\mathfrak{D}}{dt}\right)_n$$

where, in the mks system, $\mathcal{J}_n$ is measured in amperes per square meter.

118. Continuity of Current. From the fact that elementary charges are indestructible, we may conclude that an electric current is continuous and that electricity is conserved. Such is indeed the case. The proof of this statement is readily available.[1]

119. Electrical Conduction. The electric field intensity, $\mathcal{E}$, is the force that tends to move electric charges. Under the influence of such a force, a current density, $\mathcal{J}$, tends to flow in a conductor, the current density being represented by the time rate of flow of charges across an area perpendicular to the electric field intensity. The electric field intensity may then be viewed as the cause that produces or tends to produce a flow of conduction current density. There must, therefore, be some relation between $\mathcal{E}$ and $\mathcal{J}$ to express this cause-effect relation and, from what has been said about dielectrics and conductors, we might expect the results for these two cases to differ.

The relation between the electric field intensity and the current density cannot be obtained by simple reasoning alone; recourse must be made to experiment as well as to theory to obtain this relationship. Experiments show, however, that for homo-

[1] See, for example, such references as the following: R. W. P. King, "Electromagnetic Engineering," Vol. 1, McGraw-Hill, 1945; J. A. Stratton, "Electromagnetic Theory," McGraw-Hill, 1941; W. R. Smythe, "Static and Dynamic Electricity," McGraw-Hill, 1950.

geneous solids, this relationship is given by

$$\mathcal{J} = \sigma \mathcal{E}$$

where σ is a constant, characteristic of the material through which the electric field is established and through which the current flows, called the "conductivity" of the material. We may also obtain the correlation by means of the expression

$$\mathcal{E} = \rho \mathcal{J}$$

where $\rho = 1/\sigma$ is a characteristic of the material, known as its "resistivity." It should be observed that although the Greek letter sigma (σ) has been used for surface charge density as well as for conductivity, there is no connection between the two concepts expressed by the same symbol. Likewise, there is no connection between volume charge density and resistivity although both are expressed by the Greek rho (ρ). Little confusion will result because of the use of common symbols for different ideas, since the context will indicate which meaning is intended.

120. Conduction in Metals. Being an inherent property of a material to which an electric field is applied, the conductivity may be expected to be a function of the physical or chemical composition of the material, particularly with regard to those factors

Table 7. Resistivity and Temperature Coefficient of Resistivity of Common Electrical Conductors at t_0 = 20°C

Material	Resistivity, ohm-m	Temperature coefficient, A/°C
Aluminum	2.824×10^{-8}	0.0038
Antimony	41.7×10^{-8}	0.0036
Bismuth	120×10^{-8}	0.004
Brass	7×10^{-8}	0.001–0.002
Carbon	2.900×10^{-8}	−0.0005
Copper, annealed	1.724×10^{-8}	0.00393
Copper, hard-drawn	1.771×10^{-8}	0.00382
Gold	2.44×10^{-8}	0.0034
Iron	10.0×10^{-8}	0.0050
Mercury	95.783×10^{-8}	0.00089
Molybdenum	5.7×10^{-8}	0.0033
Nickel	7.8×10^{-8}	0.006
Platinum	10.0×10^{-8}	0.003
Silver	1.59×10^{-8}	0.0038
Tantalum	15.5×10^{-8}	0.00347
Thorium	40.1×10^{-8}	0.0021
Tungsten	5.6×10^{-8}	0.0045
Zinc	5.8×10^{-8}	0.00415

which determine or influence the relative freedom of motion of electrons within the material. This is, indeed, found to be true, so that the electrical conductivity or resistivity of a medium can be related to its characteristic atomic structure.

An elementary theory of metallic conduction[1] shows that the conductivity of a substance is given by the relation

$$\sigma = \frac{1}{2} \frac{l_m}{v_m} N e^2$$

where l_m is the mean free path of the electrons in the conductor, v_m is the average random velocity with which the electrons travel, N is the number of electrons in a length, l, of conductor of uniform cross section, and e is the charge of the electron. In mks units the conductivity is measured in mhos per meter. The resistivity, $\rho = 1/\sigma$, is measured in ohm-meters in the mks system of units.

The number of free electrons, N, in any given length of uniform conductor may be expected to be a function of temperature, so that the electrical properties of conductors

[1] Page, L., and N. I. Adams, "Principles of Electricity," p. 162, Van Nostrand, 1931.

are likewise a function of temperature. For metallic conductors, the resistivity increases with the temperature, but there are some materials, of which carbon is a notable example, in which the resistivity decreases with an increase of temperature. The resistivity of a homogeneous material may be expressed, as a function of temperature, by the expression

$$\rho = \rho_0[1 + A(t - t_0)]$$

where ρ_0 is the resistivity of the material at the specified or reference temperature, t_0, t is a temperature at which the resistivity is desired, and A is a constant, called the *temperature coefficient* of resistivity, typical values for which are given in Table 7.

121. Magnetic Field Produced by Electric Currents. Ampère's experiments on the magnetic intensity produced by an electric current showed that the magnetic field

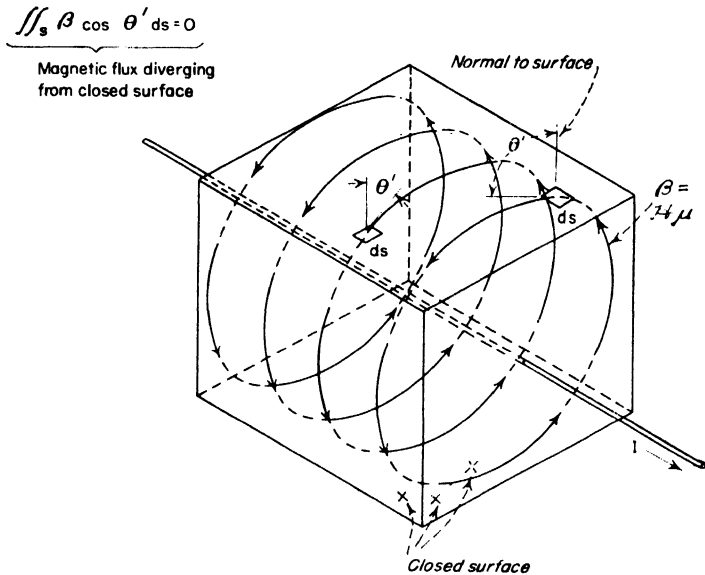



FIG. 29. Lines of magnetic intensity around a current-carrying conducor are continuous, and the total magnetic flux crossing a closed surface is zero.

intensity produced at a point is proportional to the number of conductors, N, the current, I, in each conductor, the length, l, of the current-carrying elements, and the sine of the angle, θ, between the direction of the current element and the radius vector, $r = OP$, from the current-carrying element to the point P; it is also inversely proportional to the square of the distance, r, between the current-carrying element and the point of observation. Mathematically,

$$\mathcal{H} = \frac{NIl \sin \theta}{r^2}$$

where, in mks units, the magnetic field intensity, $\mathcal{H}$, is measured in ampere-turns per meter. From the relation $\mathcal{B} = \mu\mathcal{H}$, it follows that the magnetic flux density due to the flow of current through current-carrying elements is given by

$$\mathcal{B} = \frac{\mu(NIl \sin \theta)}{r^2}$$

where $\mathcal{B}$ is measured in webers per square meter in mks units.

Lines of magnetic intensity surrounding a current-carrying conductor are represented in the diagram of Fig. 29.

122. Divergence of Magnetic Flux. For a magnetic field produced by the flow of an electric current, as well as in the magnetostatic case, the magnetic flux is given by

$$\phi = \iint_s \mathcal{B} \cos \theta \, ds$$

where θ is the angle between the direction of $\mathcal{B}$ and the normal to the surface element, ds. Over a closed surface,

$$\phi = \oiint_s \mathcal{B} \cos \theta \, ds$$

i.e., the divergence of the magnetic flux, over a closed surface, is zero. This is represented diagrammatically in Fig. 29 by the illustration showing as many lines of magnetic intensity entering as leaving a small volume containing the current-carrying element.

123. Magnetic Force Produced by Current Elements. If a conductor of N turns, through which current I flows, lies in a magnetic field whose flux density is $\mathcal{B}$ webers per square meter, the force exerted on the conductor is given by the relation

$$f = \mathcal{B}NIl \sin \theta$$

where θ is the angle between the direction of $\mathcal{B}$ and that of l.

Since a magnetic field can be produced by the flow of current through conductors, and further since a current-carrying element in a magnetic field has a force exerted upon it, it follows that a mechanical force must exist between two current-carrying conductors.

124. Force between Current-carrying Elements. It has been shown (Sec. 121) that a current-carrying circuit element establishes a magnetic field and (Sec. 123) that a force is exerted on a conductor carrying a current in a magnetic field. Therefore, if two conductor elements each carry an electric current, each will be in the magnetic field produced by the other; hence, a force will be exerted between the two. The force between these two elements is given by

$$f = K_{em}K' \left(\frac{\mu}{r^2}\right) (N_1I_1l_1 \sin \theta_1)(N_2I_2l_2 \sin \theta_2)$$

where N_1 and N_2 are the number of turns of the first and second circuit elements, I_1 and I_2 are the currents in the first and second elements, respectively, l_1 and l_2 are the lengths of the first and second conducting elements, and θ_1 and θ_2 are the angles between straight lines connecting the two current-carrying elements and the direction of the first and second elements, respectively. In mks units, f is measured in newtons, I in amperes, l in meters, and μ in henrys per meter. The factor K_{em} relates units of mechanical force with units of current, and K' is a factor, as given in Table 6, whose value depends on whether a rationalized or unrationalized system of units is employed.

125. Relation between Static and Dynamic Force Fields. Coulomb's law for electric charges and magnetic poles provides a means of relating mechanical forces with the magnitudes of electric charges and magnetic poles. On the basis of Coulomb's law for electric charges, the electrostatic system of units has been developed. Quite independently, the magnetic system of units has been derived on the basis of Coulomb's law for magnetic poles. The force between two current-carrying elements provides a means for relating the static and dynamic force fields, since currents produce magnetic effects and are but the flow of charges. It also provides a means for checking and correlating the system of units established for the electrostatic and magnetostatic cases.

Equating the forces obtained from charges, single-turn current-carrying circuit elements, and magnetic poles, the three force expressions are

$$f = K_eK' \frac{Q_1Q_2}{r^2} = K_{em}K' \frac{1}{\mu r^2} (\mu I_1l_1 \sin \theta_1)(\mu I_2l_2 \sin \theta_2) = K_mK' \frac{M_1M_2}{\mu r^2}$$

By the definition of unit current and magnetic shell, the last two expressions must be identical so that $K_{em} = K_m$. To simplify calculations, let $Q_1 = Q_2 = Q$, $M_1 = M_2 = M$, $I_1l_1 = I_2l_2 = Il$, and let $\theta_1 = \theta_2 = \theta = 90°$ so that $\sin\theta = 1$. Then, for r the same in all cases

$$K_e\left(\frac{Q^2}{\epsilon}\right) = K_m\left(\frac{M^2}{\mu}\right) = K_m\mu(Il)^2$$

For a steady current, $I = Q/t$; hence,

$$K_e\left(\frac{Q^2}{\epsilon}\right) = K_m\mu\left(\frac{Ql}{t}\right)^2$$

from which

$$\frac{K_e}{K_m}\frac{1}{\mu\epsilon} = \left(\frac{l}{t}\right)^2 = v^2$$

where the velocity v is given by l/t.

If the permittivity and permeability of free space are both taken as unity, as they are in the cgs electrostatic and electromagnetic (magnetostatic) systems of units, the ratio of units in the two systems has the dimensions of velocity squared. Further examination of the units of measurement in these two systems shows that v is the same as the velocity of light in free space to well within the limits of experimental error. For mediums other than that of free space, the actual values of μ and ϵ must be used to determine the velocity of propagation of electrical effects.

The somewhat unexpected result that the ratio of dimensions in the cgs electrostatic and magnetostatic systems has the dimensions of velocity squared comes from the fact that values for permeability and permittivity were arbitrarily assigned at a time before there was any known relation between electricity and magnetism.

126. Magnetomotive Force. The magnetomotive force acting in any closed path in a magnetic field is the line integral of the magnetizing force around the path. In any closed path, the magnetomotive force resulting from the flow of current through a path is proportional to the current that links its path. In mks units, the magnetomotive force is expressed in ampere-turns, and around any closed path has the value

$$\oint \mathcal{H}\cos\theta\, dl = K'''NI$$

where θ is the angle between the direction of $\mathcal{H}$ and that of l, N is the number of times the path links the magnetic field, I is the current in amperes, and K''' has the value given in Table 6.

127. Ampère's Circuital Law for Magnetic Intensity. In radio practice we are interested not only in conduction currents and circuit phenomena (in which current flows through filamentary circuit paths), but also in the general flow of current through space, more conveniently expressed in terms of current density. Furthermore, the current density may have displacement as well as conduction components. For the case in which current density flows through a surface, Ampère's circuital law of magnetic intensity may be stated as follows: The magnetizing force around a closed loop is equal to the total current flowing through the surface bounded by the loop. Mathematically,

$$\oint \mathcal{H}\cos\theta\, dl = \iint_S K'''\left(\mathcal{I}_c + K''\frac{d\mathfrak{D}}{dt}\right)_n ds = K'''\left(I_c + K''\frac{d\psi}{dt}\right)$$

In this case $\mathcal{I}_c$ is the conduction component of the current density and $d\mathfrak{D}/dt$ is the displacement component, while I_c is the conduction component of the total current and $d\psi/dt$ is the displacement component of current. The subscript, n, designates the normal component. In the mks system of units, current is measured in amperes, current density in amperes per square meter, and magnetic field intensity in ampere-turns per meter.

The integral form of Ampère's circuital law, above, is particularly suitable when the net over-all, or integrated effects of the resultant magnetic field and electric currents are of primary concern, as in dealing with electric circuits. When dealing with radiation phenomena, field concepts, or other matters in which the spatial distribution of electromagnetic effects is important, it is more convenient to deal with the differential form of Ampère's circuital equation in which the directional components of the current density and electric flux density are expressed in terms of the curl or rotational components of the magnetic field intensity.

In deriving the differential form of Ampère's circuital law, the resultant current flow is resolved into three orthogonal components; likewise the resulting magnetic

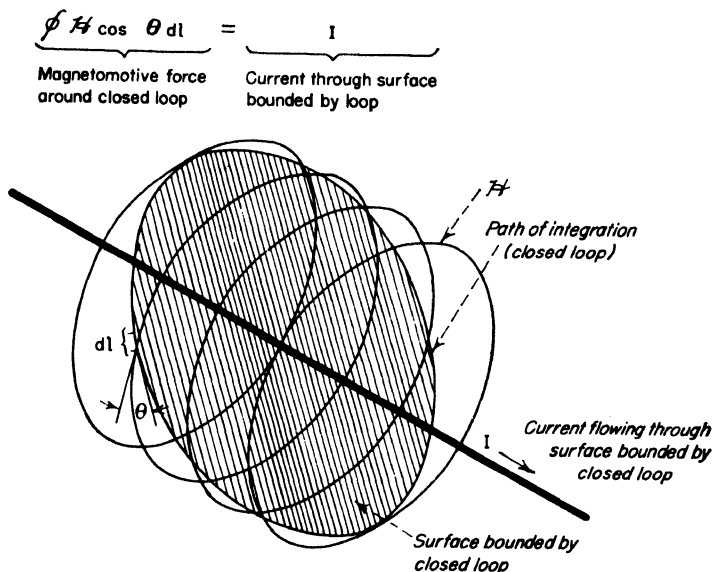



FIG. 30. Diagram illustrating the integral form of Ampère's circuital law.

field is resolved into three orthogonal components such that the circuitation of the magnetic field is in a plane normal to the direction of flow of the electric current. The resultant effect, at a point in space, is then the vector sum of the three orthogonal components.

In rectangular coordinates the differential expression for Ampère's circuital law takes the form

$$\mathbf{i}K'''\left(\mathcal{J}_x + K''\frac{d\mathfrak{D}_x}{dt}\right) + \mathbf{j}K'''\left(\mathcal{J}_y + K''\frac{d\mathfrak{D}_y}{dt}\right) + \mathbf{k}K'''\left(\mathcal{J}_z + K''\frac{d\mathfrak{D}_z}{dt}\right)$$
$$= \mathbf{i}\left(\frac{\partial\mathcal{H}_z}{\partial y} - \frac{\partial\mathcal{H}_y}{\partial z}\right) + \mathbf{j}\left(\frac{\partial\mathcal{H}_x}{\partial z} - \frac{\partial\mathcal{H}_z}{\partial x}\right) + \mathbf{k}\left(\frac{\partial\mathcal{H}_y}{\partial x} - \frac{\partial\mathcal{H}_x}{\partial y}\right)$$

The sum of the three directional components gives the relation, at a point, for Ampère's circuital law.

The diagram of Fig. 30 represents the relation existing between the direction of the flow of current, the direction of the magnetic field intensity, and the direction of a closed path, as called for by Ampère's circuital law.

128. Electromotive Force. Electromotive force (emf) is that property of a physical device which tends to produce an electric current in a circuit. It is also that property of a system which tends to alter the motion of electricity, or to maintain its motion

against resistance. *Like the difference of potential, emf is measured by the energy per unit charge imparted to the charge as it travels through a region in which the emf acts. It is evident, therefore, that the emf is not, as its name would imply, a force, but rather work per unit charge.* In the mks system of units, the unit of emf is the joule per coulomb, to which the name "volt" has been given.

One form of emf has already been encountered, that of electrostatic potential difference, which exists by virtue of the position of electric charges. Sources of electric energy, such as batteries, thermocouples, and photoelectric devices, are capable of producing an emf by chemical action or by the action of radiant energy. Another type of emf, first discovered by Faraday, results from the motion of a conductor in a magnetic field; it is called the "induced" emf, or the "electromotive force of induction."

129. Faraday's Circuital Law for Induced Emf. As a result of extensive researches, Faraday, in 1833, was able to announce his discovery of the relation governing an emf appearing across a section of a conductor having motion relative to a magnetic field in which it is situated. In the following year, Lenz clearly specified the relationships between the directions of the magnetic field, the motion of the conductor relative to the magnetic field, and the polarity or direction of the resulting emf. As amplified by Lenz, Faraday's law may be stated as follows: When the change in magnetic flux linkages is caused by the relative motion of a magnetic field and a conductor forming part of an electric circuit, an electromotive force is induced in the circuit, the magnitude of which is equal to the time rate of decrease of the flux linkages.

A flux linkage is defined as the interlinking of a line of magnetic flux with a single conductor, and is represented by the symbol Λ. Therefore, $\Lambda = N\phi$ flux linkages signifies the product of ϕ lines of magnetic flux linking N turns of a conductor or, conversely, N conductors linking the average value of the magnetic flux, ϕ. With this notation, the magnitude and direction of the emf induced in a conductor is given by

$$E = -\frac{d\Lambda}{dt} = -\frac{d(N\phi)}{dt}$$

where, in mks units, E is the induced emf in volts, N is the number of conductors (or number of turns of a single conductor closely bunched together), and ϕ is the average value of the magnetic flux in webers.

The relation given above is quite suitable for electric circuits but, when dealing with field phenomena, the equations are more convenient when expressed in terms of the electric field intensity and the magnetic flux density together with elements of length and surface. Since the drop in emf around a closed loop is given by the relation $\oint \mathcal{E} \cos \theta \, dl$ and the magnetic flux traversing the surface enclosed by the loop is given by $\iint_s \mathcal{B} \cos \theta' \, ds$, Faraday's law for electromotive induction may be written as

$$\oint \mathcal{E} \cos \theta \, dl = -\frac{\partial}{\partial t} \iint_s \mathcal{B} \cos \theta' \, ds$$

where (in mks units) $\mathcal{E}$ is the electric field intensity given in volts per meter, $\mathcal{B}$ is the magnetic flux density in webers per square meter, dl is an element of length in meters, ds is an element of area in square meters, and θ is the angle between the direction of the electric field intensity and the line element, whereas θ' is the angle between the direction of the magnetic flux density and the normal to the surface through which the flux passes.

The magnetic flux density is a continuous function of time, t, and surface, s. For such conditions it is permissible to perform the differentiation under the integral sign[1] so that Faraday's circuital law may also be written in the form

$$\oint \mathcal{E} \cos \theta \, dl = -\iint_s \frac{\partial(\mathcal{B} \cos \theta')}{\partial t} \, ds$$

where the symbolic notation is the same as before.

[1] Sokolnikoff, I. S., and E. S. Sokolnikoff, "Higher Mathematics for Engineers and Physicists," McGraw-Hill, 1941.

At a point, this field relation may be expressed in the differential form

$$\mathbf{i}\left(\frac{\partial \mathcal{E}_z}{\partial y} - \frac{\partial \mathcal{E}_y}{\partial z}\right) + \mathbf{j}\left(\frac{\partial \mathcal{E}_x}{\partial z} - \frac{\partial \mathcal{E}_z}{\partial x}\right) + \mathbf{k}\left(\frac{\partial \mathcal{E}_y}{\partial x} - \frac{\partial \mathcal{E}_x}{\partial y}\right) = -\left(\mathbf{i}\frac{\partial \mathcal{B}_x}{\partial t} + \mathbf{j}\frac{\partial \mathcal{B}_y}{\partial t} + \mathbf{k}\frac{\partial \mathcal{B}_z}{\partial t}\right)$$

where, in mks units, $\mathcal{E}$ is the electric field intensity in volts per meter, $\mathcal{B}$ is the magnetic flux density in webers per square meter, and the directional components are indicated, in rectangular coordinates, by the subscripts.

In the diagram of Fig. 31 is represented flux density, $\mathcal{B}$, flowing through the surface, s, enclosed by a looped conductor. The direction between the plane of the closed conductor and the plane normal to that of the magnetic flux is θ, and the direction or polarity of the emf induced in the loop is also indicated.

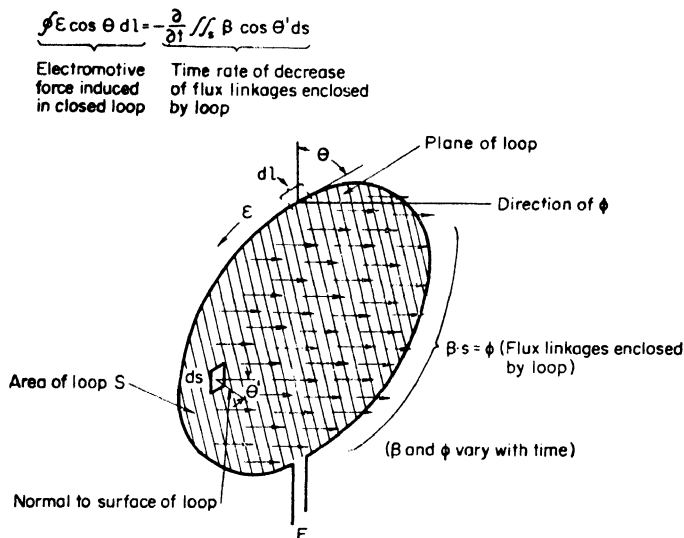



FIG. 31. Representation of integral form of Faraday's circuital law for induced emf.

130. Potential Difference, Electromotive Force, and Voltage. The terms "potential difference," "electromotive force," and "voltage" are often used loosely and indiscriminately, although there is a clear distinction between the first two of these at least.

Strictly speaking, the potential difference is the difference in electrostatic potential due to the distribution of electric charges, and the magnitude of the potential difference depends upon the magnitudes and positions of the charges. The concept is also extended, however, to the case of moving charges, *i.e.*, the case in which current flows, in which case the potential difference is interpreted, at any instant of time, to be the same as the electrostatic potential difference for the charge configuration at the instant in question. This extension of the concept of electrostatic potential difference is valid for time variations of the charges, or frequencies of current, so long as the electromagnetic field is essentially of equal magnitude, at any instant of time, for the entire region occupied by the electrical system with which it is associated. At sufficiently high frequencies, the concept is no longer applicable.

An electric field may be produced by the distribution of electric charges or, in accordance with Faraday's circuital law, it may be produced by the relative motion of a conductor and a magnetic field.

The term "voltage" is used to designate potential difference whether due to the simple or extended concepts of electrostatic potential or due to induced emf.

It is worth noting that currents are not necessarily induced. If emf is induced in a closed circuit, a conduction current will flow, but the emf will be induced whether the circuit is open or not. Conduction current cannot flow in an open circuit, however.

References

Abraham, M., and R. Becker: "Classical Theory of Electricity and Magnetism," Blackie, 1932.
Faraday, M.: "Experimental Researches in Electricity," Dutton, 1931.
Hirst, A. W.: "Electricity and Magnetism for Engineering Students," Prentice-Hall, 1937.
Kelly, H. C.: "A Textbook of Electricity and Magnetism," Wiley, 1941.
Kraus, J. D.: "Electromagnetics," McGraw-Hill, 1953.
Loeb, L. B.: "Fundamentals of Electricity and Magnetism," Wiley, 1931.
Maxwell, J. C.: "Electricity and Magnetism," Oxford, 1881.
Page, L.: "Introduction to Theoretical Physics," Van Nostrand, 1928.
——— and N. I. Adams: "Electrodynamics," Van Nostrand, 1940.
——— and ———: "Principles of Electricity," Van Nostrand, 1931.
Planck, M.: "Theory of Electricity and Magnetism," Macmillan, 1932.
Pohl, R. W.: "Physical Principles of Electricity and Magnetism," Blackie, 1930.
Poor, V. C.: "Electricity and Magnetism," Wiley, 1931.
Ramsey, A. S.: "Electricity and Magnetism," Cambridge, 1937.
Starling, S. G.: "Electricity and Magnetism for Advanced Students," Longmans, 1912.
Suydam, V. A.: "Fundamentals of Electricity and Electromagnetism," Van Nostrand, 1940.
Thomson, J. J.: "Elements of the Mathematical Theory of Electricity and Magnetism," Van Nostrand, 1921.
Thomson, W.: "Papers on Electrostatics and Magnetism," Macmillan, 1884.
Webster, A. G.: "Electricity and Magnetism," Macmillan, 1897.
Wilson, W.: "Electromagnetism and Optics," London, 1933.

THE ELECTROMAGNETIC FIELD

131. Electromagnetic Field Relations. As concerns its electrical aspects, the basis of radio communication resides in the four electromagnetic field relations known as Maxwell's equations, which have already been developed and are summarized in Table 8. They are fundamental in the sense that from them can be derived relations of circuit behavior, the wave equation, the propagation characteristics of electromagnetic waves (if enough is known of the transmission medium), and the like.

Maxwell's electromagnetic field equations should be regarded as a compact and generalized means of expressing known electrical phenomena. This set of relationships differs from most equations used by engineers in that they do not provide information in a form that is directly applicable for design purposes. In fact, the use of these equations for most problems of engineering application usually involves such difficulties of specifying boundary conditions and mathematical manipulation that the Maxwell equations have little direct usefulness to engineers. But by no means do these limitations invalidate their utility, for this set of relations can be made to yield other relationships (*e.g.*, the laws for electric circuits) applicable in special cases and sufficiently easy to manipulate as to be of great practical utility.

In Table 8 each of the four Maxwell equations specifying a fundamental electromagnetic relation is interpreted in six different ways in addition to the pictorial representation referred to by the figure numbers. The first column gives the usual name of the relationship. The second column provides a word statement of each relation, which aims to express the fundamental physical behavior without recourse to mathematical notation. The third column gives the figure number in which the approximate physical relations are pictorialized. The next columns express the result of physical observations in the classical form of integral equations which take account of the total, net, or resultant of all directional components of the field. The last columns are the equivalent integral expression in the notation of vector analysis. The integral form is the basis of circuit equations and is most suitable where the integrated electric and magnetic fields are most conveniently expressed in terms of voltage and current. Each of the field components is also expressed in the form of a differential equation, one column giving the classical form (in rectangular coordinates only) and the other giving the equivalent in vector notation. As given in Table 8, the equations are suitable for either the rationalized or unrationalized system. The differential form of the field equations is most suitable for the determination of rela-

Table 8. Alternative Statements of Maxwell's Equations

Name	Statement	Fig.	Classical form		Vector form	
			Integral	Differential (point rel.)	Integral	Differential (point rel.)
1. Gauss's theorem for electric flux density (Coulomb's law for single dielectric)	The electric flux diverging from a closed surface, ψ, is equal to K times the total density enclosed by the surface or to K times the charges enclosed by the surface enclosing the volume	28	$\psi = \oiint_s \mathcal{D} \cos \theta' \, ds = \iiint_r \rho \, dr = K'''Q$	$\frac{\partial \mathcal{D}_x}{\partial x} + \frac{\partial \mathcal{D}_y}{\partial y} + \frac{\partial \mathcal{D}_z}{\partial z} = \begin{cases} K'''\rho \\ \text{or} \\ 0 \end{cases}$	$\psi = \oiint_s \mathcal{D} \cdot n \, ds$	$\text{div}\, \mathcal{D} \equiv \nabla \cdot \mathcal{D} = \begin{cases} K'''\rho \\ \text{or} \\ 0 \end{cases}$
2. Gauss's theorem for magnetic flux density (Coulomb's law for a single magnetic medium, or law of magnetic flux density)	The magnetic flux diverging from a closed surface, ϕ, is zero. It follows that magnetic lines of force are continuous, or form closed loops	29	$\phi = \oiint_s \mathcal{B} \cos \theta' \, ds = 0$	$\frac{\partial \mathcal{B}_x}{\partial x} + \frac{\partial \mathcal{B}_y}{\partial y} + \frac{\partial \mathcal{B}_z}{\partial z} = 0$	$\phi = \oiint_s \mathcal{B} \cdot ds$	$\text{div}\, \mathcal{B} \equiv \nabla \cdot \mathcal{B} = 0$
3. Circuital law for magnetic intensity (from Ampere and Maxwell)	The magnetomotive force around a closed loop, $\mathfrak{F}$, is proportional to the total current flowing through the surface bounded by the loop	30	$\mathfrak{F} = \oint \mathcal{H} \cos \theta \, dl = \iint_s K''' \left(\mathcal{J}_c + K' \frac{\partial \mathcal{D}}{\partial t} \right)_n ds$	$\frac{\partial \mathcal{H}_z}{\partial y} - \frac{\partial \mathcal{H}_y}{\partial z} = K''' \left(\mathcal{J}_x + K'' \frac{\partial \mathcal{D}_x}{\partial t} \right)$ $\frac{\partial \mathcal{H}_x}{\partial z} - \frac{\partial \mathcal{H}_z}{\partial x} = K''' \left(\mathcal{J}_y + K'' \frac{\partial \mathcal{D}_y}{\partial t} \right)$ $\frac{\partial \mathcal{H}_y}{\partial x} - \frac{\partial \mathcal{H}_x}{\partial y} = K''' \left(\mathcal{J}_z + K'' \frac{\partial \mathcal{D}_z}{\partial t} \right)$	$\mathfrak{F} = \oint \mathcal{H} \cdot dl = \iint_s K''' \left(\mathcal{J}_c + K'' \frac{\partial \mathcal{D}}{\partial t} \right) \cdot ds$	$\text{curl}\, \mathcal{H} \equiv \nabla \times \mathcal{H} = K''' \left(\mathcal{J}_c + K'' \frac{\partial \mathcal{D}}{\partial t} \right)$
4. Circuital law for induced emf (generalized law of Faraday)	In all cases in which the flux linkages change, the algebraic value of induced emf in a closed loop, E, is equal to the time rate of decrease of flux linkages enclosed by the loop	31	$E = \oint \mathcal{E} \cos \theta \, dl = -\frac{\partial}{\partial t} \iint_s \mathcal{B} \cos \theta' \, ds = -\iint_s \frac{\partial(\mathcal{B} \cos \theta')}{\partial t} ds$	$\frac{\partial \mathcal{E}_z}{\partial y} - \frac{\partial \mathcal{E}_y}{\partial z} = -\frac{\partial \mathcal{B}_x}{\partial t}$ $\frac{\partial \mathcal{E}_x}{\partial z} - \frac{\partial \mathcal{E}_z}{\partial x} = -\frac{\partial \mathcal{B}_y}{\partial t}$ $\frac{\partial \mathcal{E}_y}{\partial x} - \frac{\partial \mathcal{E}_x}{\partial y} = -\frac{\partial \mathcal{B}_z}{\partial t}$	$E = \oint \mathcal{E} \cdot dl = -\frac{\partial}{\partial t} \iint_s \mathcal{B} \cdot ds = -\iint_s \frac{\partial \mathcal{B}}{\partial t} \cdot ds$	$\text{curl}\, \mathcal{E} \equiv \nabla \times \mathcal{H} = -\frac{\partial \mathcal{B}}{\partial t}$

tions at a point, or where it is necessary to take into account the spatial and directional distributions of the electric and magnetic effects of electromagnetic phenomena. The classical method of writing these equations is somewhat long but, for purposes of physical interpretation, has the advantage of illustrating the manner in which each directional component of the electromagnetic field is separately involved. The vector form is much more compact and has the advantage that the same compact notation is used for any orthogonal system of coordinates. A working knowledge of vector analysis is required before the physical interpretation of the equations is evident, however.

132. Maxwell's Equations for Special Cases. Maxwell's equations may be grouped into five distinct divisions, depending upon whether or not certain effects are functions of time and whether the magnitude of certain physical effects is sufficiently small to be negligible. From an engineering point of view, the separate cases are (1) the static case, (2) the steady case for stationary electric and magnetic fields, (3) the quasi-steady case for closed circuits in which, except for capacitors, the displacement component of the current is sufficiently small to be negligible, (4) the quasi-steady case for open circuits in which displacement current is no longer negligible, and (5) the general case of dynamic phenomena in which the magnitude of the displacement current is appreciable and may exceed that of conduction current. These cases are discussed below.

1. *The Static Case.* The static case is that, encountered in electrostatics and magnetostatics, for which no time variations of electric charges or magnetic poles occur. There is no flow of charge and, therefore, there can be neither conduction nor displacement currents. Except for Gauss's theorem for electric charges, the right-hand side of all the Maxwell equations reduces to zero; it may even reduce to zero for Gauss's theorem for electric charge if the medium is charge-free. The electrostatic and magnetostatic fields have independent laws (the divergence equations deduced from Coulomb's law) even though the mathematical forms of these laws are similar. Energy is associated with the electrostatic field and also, separately, with the magnetostatic field. The electrostatic and magnetostatic fields become superimposed upon one another without influencing each other in the slightest degree; hence, the complete segregation of the two kinds of fields and the impossibility of determining the connection between electricity and magnetism for the static case.

Although the static case has been important in the development of the classical theory of electricity, it can obviously have no direct application in conveying electrical communication, for the essence of electrical communication is the time variation of some electrical quantity in accordance with the information to be transmitted.

2. *Steady Case for Stationary Electromagnetic Field.* The steady case is that, encountered in d-c circuits and systems, for which there is a constant and steady flow of electric charges to produce a conduction (or convection) current, but no displacement current is possible. All of Maxwell's equations may be applied here except that expressing Faraday's circuital law of induced emf which, of course, depends upon a time-varying magnetic field. Electric and magnetic effects are now no longer independent of one another. The electric field in a conductor establishes a streamline of flow of electric charges which, in turn, gives rise to a magnetic field that is constant with respect to time and does not react to modify the electric field. A continual transformation of electric energy into heat takes place. Lumped constant circuit elements can be rigorously defined for this case; in fact, it is the only case for which the circuit parameters are rigorously defined. Yet, in the true steady case, *i.e.*, free from transients, only the resistance parameter is encountered in circuit behavior; inductance and capacitance determine circuit behavior only for transient conditions.

The steady case—the d-c or stationary electromagnetic field case—has no direct active role in electrical communication that requires time variations of some electrical quantity. Nevertheless, the steady case is of considerable indirect interest to engineers because certain important concepts (*e.g.*, lumped circuit parameters) are rigorously defined only for this case. Moreover, steady voltages and currents are often

required in order that certain pieces of radio equipment (*e.g.*, electron tubes) may operate properly in their communication application.

3. *Quasi-steady Case for Closed Circuits.* This is probably the best known of the five cases since it is the most frequently encountered in electrical engineering. For this case the conduction current density, the magnetic flux linkages, and the induced emf are all single-valued functions of time and (except in capacitors) the displacement current, although not zero, is negligible compared to the conduction (or convection) component and hence may be ignored or neglected. For this case consideration is restricted to electrical processes occurring so slowly that, at any instant of time, the entire electromagnetic field surrounding the electrical system may be regarded as uniform. The electric and magnetic effects for this case are mutually related to the extent that a varying current produces a varying magnetic field which is capable of inducing a varying emf in a neighboring conductor which establishes a varying current if the conductor is closed. The fact that the displacement current is negligible precludes the possibility of radiation, or the mutual self-sustenance of electric and magnetic field intensities in dielectrics. All Maxwell equations have interpretation, but (except in capacitors) displacement currents are negligibly small and, therefore, are omitted from consideration. At any instant of time the current at all points in a (series) circuit is the same. The magnitude of the displacement current in capacitors is equal to the magnitude of the conduction current in a series circuit. A very convenient and highly useful simplification of Maxwell's equation results from neglecting the displacement current.

The very great advantage obtained for this case is that the number of variables determining the state of the electromagnetic field is greatly reduced. In fact, for most problems in this case, it is possible to omit consideration of space variables and deal only with time variations of the voltage and current—the integrated effects of the electric and magnetic field components. In the case of linear conductors the analysis can be still further simplified, for the current is determined by the voltage; hence, the entire phenomena may be expressed in terms of a single varying time function—either the current or the voltage. Usually the applied voltage is known, and if the current is then determined as a function of time, the problem of electrical behavior for this case is considered to be solved.

As in the steady case, the electric field intensity establishes a streamline of flow (tangential along the conductors) of electric charges of varying magnitude producing a varying magnetic field about the conductor.

For this case, lumped circuit parameters are no longer rigorously defined because displacement current is not taken into account. Therefore, the use of lumped circuit parameters leads to approximate rather than exact solutions of the electric circuit behavior. The errors involved in such approximations are often sufficiently small to be completely negligible for engineering (and even for a good deal of scientific) work. The magnitude of the displacement current relative to that of the conduction current—and, hence, the error due to circuit analysis based on steady-case derivations—increases with frequency. Such error becomes appreciable (usually for frequencies above 50 to 100 Mc) for those cases in which the dimensions of the electrical system are comparable to a wavelength of oscillation or (what is the same thing) for those cases for which the instantaneous electromagnetic field can no longer be regarded as uniform throughout the space occupied by the electrical system. Distributed circuit parameters, *i.e.*, those circuit parameters whose properties are spatially distributed throughout the region of the electrical system instead of being considered as lumped at a point in space, may often be used in place of lumped circuit parameters to obtain solutions of electric circuits (such as transmission lines) in which at least one dimension is an appreciable portion of a wavelength or more.

4. *Quasi-steady Case for Open Circuits.* This case is identical to that above with the following exceptions: Because displacement currents are no longer of negligible magnitude, the current at all points of an open circuit is no longer the same as in the case for closed circuits. The flow of energy is no longer tangential along conductors

but has a perpendicular component as well. The electric energy of the field may no longer be neglected in comparison with the magnetic energy, and the magnetic energy can no longer be expressed accurately by conditions that apply for stationary currents. Radiation phenomena begin to become important, for energy may leave the open circuit and be propagated into space. These conditions are a result of the fact that the displacement current can no longer be regarded as negligible, although for this case it is not yet regarded as exceeding the magnitude of the conduction current.

5. *General Case for Dynamic Phenomena.* For this case, no restrictions are placed on Maxwell's equations. It is usually exceedingly difficult to apply the general equations to engineering problems, and this practical difficulty impedes the application of the general equations. For this case, the displacement current has appreciable magnitude. All four Maxwell equations (and all terms in each equation) are significant and must be used in their complete form. Circuit concepts are, generally, no longer valid. For this case, the electromagnetic field can no longer be regarded as uniform throughout the spatial distribution of the electrical system at any instant of time. Consequently the general case must be employed when very high frequencies are encountered or when the dimensions of the electrical system are comparable to a wavelength of oscillation.

Because displacement currents are appreciable, there is a mutual and self-sustained reaction between the electric and magnetic components of the electromagnetic field that leads to the phenomena of radiation.

133. Circuit and Field Phenomena. For convenience in radio engineering practice, Maxwell's equations may be divided into two broad categories in which the quasi-steady case for closed circuits and the general case are the most frequently employed.

Table 9. Comparison of Electromagnetic Field and Electric Circuit Concepts*

Electromagnetic Field Concept	Electric Circuit Concept
Electric field intensity at a point in the medium	Electric voltage in branches of an electric circuit
Magnetic field intensity at a point in the medium	Electric current flowing through an electric circuit branch
Resistivity, or resistance per unit cube, of the medium	Resistance in a branch of the circuit
Permeability of the medium	Inductance of a branch of the circuit
Permittivity of the medium	Capacitance of a branch of the circuit
Energy stored in the electric field for unit volume of the medium	Energy stored in the dielectric of a capacitor
Energy stored in the magnetic field for unit volume of the medium	Energy stored in the magnetic field about an inductor
Energy dissipated in unit volume of the medium	Energy dissipated as heat in a conductor or resistor

* This table is a slight modification of that appearing in J. G. Brainerd, G. Koehler, H. J. Reich, and L. F. Woodruff, "Ultra-high-frequency Techniques," p. 371, Van Nostrand, 1942.

Upon the first of these cases is built the highly convenient theory using lumped circuit parameters, which are usually but not necessarily considered to be linear. The electromagnetic wave theory (including radiation phenomena and the behavior of electrical systems at very high frequencies) must necessarily be built upon the more general case of Maxwell's equations. The points of view in these two cases differ appreciably, as does also the emphasis with respect to spatial considerations. The approximately analogous concepts for these two cases are given in Table 9, which is intended to draw attention to the change in emphasis required for the two cases rather than to establish rigorous parallelisms.

The relations between circuit and electromagnetic field theory have been ably stated by Carson[1] in the following words:

> The equations of electric circuit theory in their present form are essentially a generalization of the observations of Ohm, Faraday, Henry, Kirchhoff, and others, and their development preceded the electromagnetic theory of Maxwell and Lorentz. Naturally, in view

[1] Carson, John R., Electromagnetic Theory and the Foundations of Electric Circuit Theory, *Bell System Tech. J.*, **6** (1), 1, January, 1927.

of its early development, circuit theory embodies approximations, the precision of which cannot be determined from the observations on which it is based. For example, circuit theory explicitly ignores the finite velocity of propagation of electromagnetic disturbances, and hence the phenomena of radiation. Again it involves the assumption that the network can be represented by a finite number of coordinates and, thus, that it constitutes a rigid dynamic system. The rigorous equations of electromagnetic theory formulate the relations between current and charge densities and the accompanying fields. Circuit theory, on the other hand, expresses approximate relations between total currents and impressed electromotive forces.

References

Abraham, M., and R. Becker: "The Classical Theory of Electricity and Magnetism," Blackie, 1932.
Atwood, S. S.: "Electric and Magnetic Fields," Wiley, 1949.
Boast, W. B.: "Principles of Electric and Magnetic Fields," Harper, 1948.
Brainerd, J. G., G. Koehler, H. J. Reich, and L. F. Woodruff: "Ultra-high-frequency Techniques," Van Nostrand, 1942.
Carson, J. R.: Electromagnetic Theory and the Foundations of Electric Circuit Theory, *Bell System Tech. J.*, **6** (1), 1, January, 1927.
Duff, A. W., and S. J. Plimpton: "Elements of Electromagnetic Theory," Blakiston, 1940.
Frank, N. H.: "Introduction to Electricity and Optics," McGraw-Hill, 1950.
Hague, B.: "Electromagnetic Problems in Electrical Engineering," Oxford, 1932.
Harnwell, G. P.: "Principles of Electricity and Magnetism," McGraw-Hill, 1949.
Heaviside, O.: "Electromagnetic Theory," 3 vols., London, 1893.
Hund, A.: "Phenomena in High-frequency Systems," McGraw-Hill, 1936.
King, R. W. P.: "Electromagnetic Engineering," Vol. 1, McGraw-Hill, 1945.
Kraus, J. D.: "Electromagnetics," McGraw-Hill, 1953.
Maxwell, J. C.: "Treatise on Electricity and Magnetism," Oxford, 1904.
Pender, H., and K. McIlwain: "High Frequency Alternating Currents," Wiley, 1931.
——— and S. R. Warren: "Electric Circuits and Fields," McGraw-Hill, 1943.
Pierce, G. W.: "Electric Oscillations and Electric Waves," McGraw-Hill, 1920.
Ramo, S., and J. R. Whinnery: "Fields and Waves in Modern Radio," Wiley, 1944.
Rogers, W. E.: "Introduction to Electric Fields," McGraw-Hill, 1954.
Sarbacher, R. A., and W. Edson: "Hyper and Ultra High Frequency Engineering," Wiley, 1943.
Schelkunoff, S. A.: "Electromagnetic Waves," Van Nostrand, 1943.
Selgin, P. J.: "Electrical Transmission in Steady State," McGraw-Hill, 1946.
Skilling, H. H.: "Fundamentals of Electric Waves," Wiley, 1942.
Slater, J. C.: "Microwave Transmission," McGraw-Hill, 1942.
Smythe, W. R.: "Static and Dynamic Electricity," McGraw-Hill, 1950.
Stratton, J. A.: "Electromagnetic Theory," McGraw-Hill, 1941.

ELECTRIC CIRCUITS

134. Electric Circuit Concepts. An electric circuit is a path or a group of interconnected paths capable of carrying conduction electric currents. A closed electric circuit is a continuous path in the form of a loop or group of interconnected loops in which each loop is capable of carrying an electric current.

An assembly of lumped circuit elements is an electric circuit in which dissipation of energy and the storage of electric and magnetic energy may be considered as concentrated separately in each of three different circuit elements of physically small size. The physical size of lumped circuit elements is sufficiently small (theoretically infinitesimal) that, at any instant of time and over the region of space occupied by the circuit, the electric and magnetic fields are uniform. This condition is rigorously true for the steady (d-c) case and very nearly true if time variations of the electric and magnetic field are sufficiently slow. A circuit element is a physical device (usually with two terminals) exhibiting the electrical property of (1) energy dissipation or conversion of electrical energy into heat in a resistor, (2) storage of electric energy in a capacitor (or condenser), or (3) storage of magnetic energy in an inductor, or coil. These three circuit elements dissipate or store energy, and a drop of potential difference exists across them when an electric current flows through them. A fourth kind of circuit element, called a "source" of electric energy, transforms chemical, mechanical, or other energy into electric energy and has a rise of potential across its terminals.

A linear circuit is an electric circuit in which a linear differential equation precisely expresses the relations between the magnitudes of the instantaneous current and those

of the rises and drops of potential difference across the circuit elements. A linear circuit is composed of circuit elements whose parameters (*i.e.*, properties for storing or dissipating energy) are independent of the magnitude of the voltage across, or the current through, the elements. Those circuit elements whose properties are dependent upon the voltage or current are termed "nonlinear" circuit elements.

Lumped or concentrated circuit elements provide a convenient means for dealing with the integrated effects of electrical phenomena since (except in the case of capacitors in which electric energy is stored) only conduction current is important. The magnitude of the displacement current in capacitors is equal to the conduction current in the leads to the capacitor; hence, circuit phenomena are completely expressible in terms of the conduction current.

In developing the concepts of lumped circuit elements, all electrical effects, including the electric and magnetic fields, taking place are limited to the immediate vicinity of the electrical system or circuit. A source of electric energy (rise of potential difference) establishes electric field intensity along the physical configuration of the circuit elements, and all currents flow along the circuit conductors as a streamline, in the same direction as that of the electric field intensity, or the drop in potential difference.[1] The geometric configuration of the circuit, therefore, determines the space geometry of the system. Accordingly, it is possible to dispense with considerations of space coordinates in lumped circuit phenomena and consider only the time variations of the integrated electric field intensity and current density, *i.e.*, the time variations of the potential differences and the current. The concept of lumped circuit elements permits the solution—even if only approximate—of many problems that are beyond our present ability to solve by means of the more rigorous method employing electromagnetic field relations.

Strictly speaking, lumped circuit concepts apply only to the case in which the electric field intensity and current density are invariant with time, *i.e.*, to the steady electromagnetic field or the d-c case. As a practical matter, with suitable modifications or extensions, the steady-case concepts are applicable to those cases in which time variations of $\mathcal{E}$ and $\mathcal{J}$ (or E and I) are present, provided the dimensions of the circuit are sufficiently small so that, at any instant of time, the electric and magnetic fields are essentially uniform throughout the region occupied by the circuit and provided that (except in capacitors) displacement currents are negligible and, therefore, radiation is inappreciable.

The idea of lumped circuit elements applies to conducting circuits composed of filamentary conductors through which the current flows and along which there are differences of potential or of electric field intensity. An exception to this statement must be made for capacitors, for the capacitance parameter necessarily involves a volume, across opposite faces of which exists a difference of potential.

Circuit phenomena deal with the integrated effects of $\mathcal{E}$ and $\mathcal{J}$; hence, these two quantities may be replaced by potential difference (E) and current (I), respectively. In developing the concept of lumped circuit constants, use is made of the relations

$$\mathcal{B} = \mu\mathcal{H} \qquad \mathcal{J} = \sigma\mathcal{E}$$
$$\mathcal{D} = \epsilon\mathcal{E} \qquad \mathcal{E} = \rho\mathcal{J}$$

where μ, ϵ, σ, and ρ may vary with temperature but, for linear circuit elements, are invariant with respect to potential differences across, or current through, the circuit element. In practice this restriction does not impose objectionable limitations, except that analyses based on the assumption of constant permeability imply that corrections or modifications of the linear circuit theory are required in dealing with circuit elements employing ferromagnetic materials. The principle of superposition (Sec. 146) applies only to systems composed of linear circuit elements.

Three conditions exist for formulating the circuit parameters of lumped circuit elements. In each case the property of storing or dissipating electrical energy is

[1] This statement applies to the conventional direction of the current; the electron flow is opposite to that stated here.

represented by a different circuit element. In conductors the electric fields are accompanied by pure conduction currents and the dissipation of electrical energy in the form of heat. These are characteristic properties of resistors. The electric fields in dielectrics are accompanied by displacement currents and the storage of electrical (potential) energy within the dielectric. This is a characteristic property of capacitors. Magnetic fields are associated with the conduction of currents and are accompanied by the storage of magnetic (kinetic) energy. This is a characteristic property of inductors.

The circuit properties of resistance, capacitance, and inductance as employed in electrical analysis are idealized concepts. As such they represent hypothetical and limiting cases that can be approached but never fully attained in the actual embodiment of mathematical ideas as physical pieces of equipment. No physical circuit element embodies purely and completely only that circuit property for which it is named. All embody more or less perfectly that circuit property for which they are named, but they also embody the remaining two circuit properties. In well-designed circuit elements the extent to which the element conforms to the mathematical ideal circuit property is a matter of design and application of the element. The extraneous circuit properties can usually be made so small as to be of minor importance, except where the frequency of variation of the voltage and current is high. In any case, the actual behavior of any physical circuit element can be represented to any desired degree of approximation or precision by suitable combinations of idealized circuit parameters.

The idealized circuit properties discussed in the following sections are those rigorously defined for the steady case. This conventional point of view fails to take in account certain matters, such as the depth of penetration and phase differences, which become important at very high frequencies and require modifications of the treatment given here for extension to high frequencies. For treatment of these more advanced topics, the reader is referred to other chapters of this handbook or to references giving more specialized treatments of this topic.[1]

135. Resistance and Conductance. Resistance is the (scalar) property of an electric circuit, or of any body that may be used as a part of an electric circuit, which determines for a given current the rate at which electric energy is converted into heat or radiant energy and which has a value such that the product of the resistance and the square of the current gives the rate of energy conversion. In the general case, resistance is a function of the current, but the term is most commonly used in connection with circuits in which the resistance is independent of the current.

Conductance is the property of an electric circuit, or of a body that may be used as a part of an electric circuit, which determines, for a given emf in the circuit or for a given potential difference between the terminals of a part of a circuit, the rate at which electric energy is converted into heat or radiant energy, and which has a value such that the product of the conductance and the square of the emf, or potential difference, gives the rate of energy conversion. In the general case, conductance is a function of the potential difference, but the term is most often used in connection with circuits in which the conductance is independent of the potential difference.

If an electric field is applied to the ends of a conductor of uniform cross section made of homogeneous material, a potential difference will exist across the conductor and a current will flow through it, the charges (electrons) flowing along the streamline of the electric field ($\theta = 0$), since they are assumed not to flow through the dielectric that surrounds the conductor. For the steady case, the current flow will be of uniform current density over each part of the cross-sectional area,[2] so that if s is the area normal to the direction of the conductor, $\mathcal{J} = I/s$. If ρ is the resistivity of the homogeneous material of which the conductor is composed so that $\mathcal{E} = \rho\mathcal{J}$, the potential

[1] These topics are treated in such works as G. P. Harnwell, "Principles of Electricity and Electromagnetism," McGraw-Hill, 1938; A. Hund, "Phenomena in High-frequency Systems," McGraw-Hill, 1936; L. Page, "Principles of Electricity," Van Nostrand, 1931; S. Ramo and J. R. Whinnery, "Fields and Waves in Modern Radio," Wiley, 1944.

[2] This condition is not fulfilled for high frequencies.

difference between points A and B of the filamentary conductor may be expressed as

$$E_{AB} = \int_B^A \mathcal{E} \cos \theta \, dl = \int_B^A \rho \mathcal{J} \cos \theta \, dl = \int_B^A \rho \frac{I}{s} \cos \theta \, dl$$
$$= \int_B^A \rho \frac{I}{s} dl = \rho \frac{l_{AB} I}{s} = RI$$

where $\theta = 0$, and $R = \rho l/s$ is the resistance of the conductor of area s between the points A and B. From a physical point of view, the resistance of a conductor may be regarded as the opposition to the flow of the conduction current through a conductor of specified cross-sectional area, specified resistivity, and specified length. The resistivity, on the other hand, is an inherent property of the material of which the resistor is made and is in no way related to the dimensions of the conductor or resistor.

The resistance of a conductor of uniform cross section and made of homogeneous material may be expressed in terms of the electric field intensity and the current density, or in terms of their integrated effects, voltage, and current, or, alternatively, in terms of the conductor dimensions and resistivity of the material of which the resistor is made. From the foregoing results we have the relation

$$R = \frac{E}{I} = \frac{\int \mathcal{E} \cos (\mathcal{E}, l) \, dl}{\int \cos (\mathcal{J}, n) \, ds} = \frac{\int \mathcal{E} \cdot dl}{\int \sigma \mathcal{E} \cdot ds} = \frac{\mathcal{E} \int dl}{\sigma \mathcal{E} \int ds} = \frac{l}{\sigma s} = \rho \frac{l}{s}$$

where $\int \mathcal{E} \cdot dl$ and $\int \mathcal{J} \cdot ds$ are the field equations for the steady-state-case electric field intensity and current density, respectively. Although the resistance has been specified in terms of the rather special case of a conductor of uniform cross section composed of homogeneous material, this procedure is largely a matter of convenience; the ratio of the two field equations specifies the resistance for any material of whatever physical form the resistor may take.

In the derivation of the constant resistance (above), it was assumed that the material was homogeneous and, therefore, that the resistivity and conductivity were constant. For many materials, these characteristic properties of material are not constant but depend upon the current through, or the potential difference across, the conductor. For such cases, E and I are not linearly related, since the resistance has a unique value for different magnitudes of voltage across or current through it. Since the resistance is no longer constant, it is customary to define the nonlinear resistance as the limiting ratio of the voltage increment to the resulting current increment, as the latter approaches zero. The variational or differential resistance thus defined is usually designated by the letter r and is given mathematically by the relation

$$r = \lim_{\Delta I \to 0} \frac{\Delta E}{\Delta I}$$

where the voltage increment across the resistance is ΔE and the corresponding current increment is ΔI.

The reciprocals of the ohmic resistance, R, and of the variational resistance, r, are sometimes more useful electrical quantities than R and r for circuit calculations. These reciprocal quantities are called the conductance (designated by $G = 1/R$) and the "variational conductance" ($g = 1/r$) and are defined by the equations

$$G = \frac{I}{E} = \frac{\int \int \mathcal{J} \cos (\mathcal{J}, n) \, ds}{\int \mathcal{E} \cos (\mathcal{E}, l) \, dl} = \frac{\mathcal{J} \int \int ds}{\rho \mathcal{J} \int dl} = \frac{s}{\rho l} = \frac{1}{R}$$

for linear conductors, or

$$g = \lim_{\Delta E \to 0} \frac{\Delta I}{\Delta E}$$

for nonlinear conductors.

Resistance is measured in volts per ampere or in ohms in the mks system, and conductance is measured in reciprocal ohms or in mhos.

The linear resistance and linear conductance express an electrical characteristic of those materials or devices obeying Ohm's law. This includes most good conductors encountered in radio practice, particularly those of metals. It does not include conductors composed of carbon, nonhomogeneous materials, or those operating over an appreciable temperature range. Ohm's law does not usually apply for semiconductors. The variational resistance is commonly employed in such nonlinear devices as electron tubes, electric arcs, copper oxide, selenium, or other rectifiers, thyrite nonlinear conductors, and temperature-sensitive conductors. The resistivity of materials commonly employed in the construction of resistors for radio purposes is given in Table 7, Sec. 120.

136. Capacitance and Elastance. Capacitance is that property of a system of conductors and dielectrics which permits the storage of electricity when a potential difference exists between the conductors. The magnitude of the capacitance is

Table 10. Relative Permittivity and Breakdown Voltage of Typical Dielectrics Used in Radio Engineering

Material	Relative permittivity	Breakdown voltage, kv/cm
Ebonite	2.80	300–1,100
Glass	5.4–9.9	300 1,500
Gutta-percha	3.3–4.9	80–200
Mica	5.6–5.8	1,500–2,000
Paraffin	2.20	350
Porcelain	5.73	
Quartz, fused	3.9	
Rubber	......	150–500
Sulfur (amorphous)	3.9	
Wood	4.2–7.8	

expressed as the ratio of the quantity of electricity on the conducting elements to the potential difference between them. If Q is the charge stored and E is the difference of potential between the conductors, the capacitance is defined to be

$$C = \frac{Q}{E}$$

If there are two conductors in a homogeneous, isotropic medium between which an electric field intensity, $\mathcal{E}$, produces a difference of potential $E = \int \mathcal{E} \cos \theta \, dl$ and if the electric flux between the two conductors is $\psi = \iint \mathfrak{D} \cos \theta' \, ds$ so that the charge on the conductor is $Q = K'\psi$, the capacitance is defined to be

$$C = \frac{Q}{E} = \frac{K'\psi}{E} = \frac{K' \iint_s \mathfrak{D} \cos \theta' \, ds}{\int \mathcal{E} \cos \theta \, dl} = \frac{K'\epsilon\mathcal{E} \iint_s ds}{\mathcal{E} \int dl} = K' \frac{\epsilon s}{l}$$

where ϵ is the permittivity of the dielectric material, s is the surface area of the conducting plates bounding the dielectric, and l is the distance between the conducting surfaces. It is interesting to observe that this expression shows the physical dimensional factors upon which the capacitance is based. The permittivity and breakdown voltage of some representative dielectrics are given in Table 10.

The expression above gives the capacitance as defined for static conditions since it has been derived on the assumptions that $\mathfrak{D}$ and $\mathcal{E}$ were time-invariant. Exactly the same expressions are obtained if $\mathfrak{D}$ and $\mathcal{E}$ are functions of time, but in this case the equation obtained for capacitance may now be expressed for the dynamic cases in

terms of varying values of potential difference and displacement current. For the dynamic case the expression for capacitance becomes

$$C = \frac{[d(K'\psi)/dt]}{dE/dt} = \frac{K'(d\psi/dt)}{dE/dt} = \frac{K' \iint_s (d\mathfrak{D}/dt)\, ds}{dE/dt} = \frac{I_d}{dE/dt} = I_d \frac{dt}{dE}$$

from which $I_d = C(dE/dt)$; hence, by integrating, the voltage-current relation for any wave form becomes

$$E = \int \frac{I_d}{C}\, dt$$

For electric circuits, the displacement current in capacitors is the same as the conduction current in the leads to the capacitor so that $I_c = I_d$; therefore, we may express the displacement current in terms of the conduction current.

It is easy to show that the static and dynamic equations for the capacitance, as derived above, are identical. From the static definition, $Q = CE$, and since $Q = \int I\, dt$ it follows that $CE = \int I\, dt$. Hence, by differentiating, $I = C(dE/dt)$, which is the kinetic definition given above.

The capacitance, as defined above, is the total capacitance, since the total electric flux originating on one surface or conducting plate is assumed to flow to the opposite conducting plate or surface. If there are a number of conductors separated by a dielectric and so arranged that the flux from conductor A flows to other conductors, B, C, . . . , N, then the total capacitance of conductor A with respect to all other conductors is given by the expression

$$C_A = \frac{K' \iint_s \mathfrak{D}\cdot ds}{\int_B^A \mathcal{E}\cdot dl} + \frac{K' \iint_s \mathfrak{D}\cdot ds}{\int_C^A \mathcal{E}\cdot dl} + \cdots + \frac{K' \iint_s \mathfrak{D}\cdot ds}{\int_N^A \mathcal{E}\cdot dl}$$

$$= C_{AB} + C_{AC} + \cdots + C_{AN}$$

In devices intended to exhibit primarily the property of capacitance (capacitors or condensers), the capacitance is the total capacitance between two plates or other conductors separated by air, oil, mica, or other suitable gaseous, liquid, or solid dielectric. Partial capacitances are encountered when more than two conductors are separated by a dielectric, as when circuits are coupled through capacitors. It is also encountered in the capacitance between the electrodes of a multielement electron tube, and sometimes these interelectrode capacitances are responsible for undesired behavior of the tube, even though the capacitances are fairly small.

Sometimes it is more convenient to employ another circuit concept, the *elastance*, which is the reciprocal of the capacitance. The elastance, represented by the letter S, may be either the total or the partial elastance, corresponding to the total or the partial capacitance. The total elastance is defined by the electrostatic relation

$$S = \frac{1}{C} = \frac{E}{Q} = \frac{K''E}{\psi} = \frac{K''\int \mathcal{E} \cos\theta\, dl}{\iint_s \mathfrak{D} \cos\theta\, ds} = \frac{K''\int \mathcal{E}\cdot dl}{\iint \mathfrak{D}\cdot ds} = \frac{K''\mathcal{E}\int dl}{\epsilon\mathcal{E}\iint ds} = \frac{K''l}{\epsilon s}$$

or, by the kinetic or dynamic relation,

$$S = \frac{1}{C} = \frac{E}{\int I\, dt}$$

Capacitance is measured in coulombs per volt, to which the name "farad" has been given. The farad is a large value of capacitance for engineering purposes, so that subdivisions of it are more commonly employed. A capacitor has a capacitance of 1 farad when a charge of 1 coulomb is sustained by a difference of potential of 1 volt

between the conductors. The microfarad (10^{-6} farad, abbreviated μf) and the micro-microfarad (10^{-12} farad, abbreviated $\mu\mu f$) are units of capacitance commonly used in radio engineering. Sometimes the abbreviations mf and mmf are used for μf and $\mu\mu f$, but such practice is not in conformity with the usually accepted abbreviations for metric prefixes (see Sec. 170).

The unit of elastance is the reciprocal farad or *daraf*, which is the elastance represented by 1 volt between two conductors whose charge is 1 coulomb.

137. Inductance. Inductance is the (scalar) property of an electric circuit, or of two neighboring circuits, which determines the emf induced in one of the circuits by a change of current in either one.

Self-inductance is the property of an electric circuit that determines, for a given rate of change of current in the circuit, the emf induced in the same circuit, the induced voltage, E, and the self-inductance, L, being related to the current flowing in the same circuit by the relation

$$E_1 = -L\frac{dI_1}{dt}$$

where L is the coefficient of self-inductance.

Mutual inductance is the common property of two associated electric circuits that determines, for a given rate of change of current in one of the circuits, the emf induced in the other. Thus, if the mutual inductance, M, is positive,

$$E_1 = -M\frac{dI_2}{dt}$$

$$E_2 = -M\frac{dI_1}{dt}$$

where E_1 and I_1 are the voltage induced and the current, respectively, in circuit 1, and E_2 and I_2 are the voltage induced and the current in circuit 2, and M is the coefficient of mutual inductance.

138. Self-inductance. The relation between induced voltage, current, and self-inductance may be obtained directly from the definition of self-inductance and induced emf. The inductance is defined as the ratio of the flux linkages to the current, or $L = \Lambda/I = N\phi/I$, whereas the induced emf is equal to the time rate of decrease of flux linkages, or $E = -(d\Lambda/dt) = -[d(N\phi)/dt]$. From the first of these relations we have $N\phi = LI$ and, combining this with the second relation, there results

$$E = -\frac{d(N\phi)}{dt} = -\frac{d(LI)}{dt} = -\left(L\frac{dI}{dt} + I\frac{dL}{dt}\right)$$

If the inductance does not vary with time, the familiar result is obtained

$$E = -L\frac{dI}{dt}$$

The general expression for induced emf may be used to derive the expression for the inductance in a way that gives a somewhat better picture of the physical action taking place. In a multiturn coil of N turns in a magnetic field whose average flux is ϕ webers not all the turns cut all the flux lines even when the turns are close together. If K is the fraction of the flux linkages out of the total maximum number that is possible, the number of actual flux linkages is $KN\phi$. For most coils in radio work (except those in which mutual coupling between portions of the coil is adjustable or variable), K and N are constant. For this special but very common case, the induced voltage is

$$E = -KN\frac{d\phi}{dt}$$

The magnetic flux may be produced by current flowing through the windings of the coil, and in this case the magnetic flux produced is given by $\phi = NI$. Hence the voltage induced in the coil by current flowing through it is given by the relation

$$E = -KN^2\frac{dI}{dt} = -L\frac{dI}{dt}$$

from which $L = KN^2$. From this point of view the self-inductance is an electrical quantity whose magnitude depends upon the square of the number of turns of the coil and also upon the fraction of the total flux linkages that actually exist for the coil in question.

The self-inductance can also be derived from field relations. Provided it is not in a region of ferromagnetic materials, *i.e.*, so long as μ is constant, the inductance is the constant ratio of the magnetic flux linkages to the current flowing. Hence, from the field relations for the steady case, the self-inductance may be defined by the relation

$$L = \frac{\phi}{I} = \frac{\iint_s \mathcal{B}\cos\theta'\, ds}{K'''\int \mathcal{H}\cos\theta\, dl}$$

For a solenoid of N turns (the solenoid being infinitely long so that end effects may be neglected) the number of flux linkages is $KN\phi$, and the self-inductance of the multi-turn coil is then

$$L = \frac{KN\iint_s \mathcal{B}\cos\theta'\, ds}{K'''\int \mathcal{H}\cos\theta\, dl}$$

For a coil whose length, l, is very much greater than its diameter, the value of $\mathcal{B}$ at the central axis is $\mathcal{B} = \mu NI/l$, so that for such a coil the inductance is given by

$$L = \frac{KN\iint_s \frac{\mu NI}{l}\cdot ds}{K'''\int \mathcal{B}\cdot dl} = \frac{KN^2\mu Is}{K'''Il} = \frac{KN^2\mu s}{K'''l}$$

In the form here given, this formula is not useful for computing the self-inductance of solenoids unless the individual factors can be evaluated. Nevertheless, this result is useful for illustrating the physical factors and dimensions upon which the self-inductance of a solenoid depends.

Self-inductance is measured in henrys or convenient decimal submultiples of the henry, such as the millihenry or the microhenry.

A few authors sometimes make use of the reciprocal inductance, designated by the symbol Γ. The reciprocal inductance is, of course, defined to be $\Gamma = I/\phi$ and is measured in reciprocal henrys or *yrnehs*.

139. Mutual Inductance. The determination of mutual inductance requires consideration of two adjacent loops. If the current flows through one coil or loop, it establishes a magnetic field around its own turns, but it also establishes a magnetic field around the adjacent loop or coil in which it induces an emf. Conversely, the current through the second coil sets up a magnetic field which induces an emf in the first coil as well as in its own turns. Let I_1 and I_2 be the current in coils 1 and 2, respectively. Let the magnetic flux linking circuit 2 as a result of current in circuit 1 be ϕ_{21}, and let ϕ_{12} represent the magnetic flux linking circuit 1 as a result of current in coil 2. Then, if K is the fraction of the number of lines of magnetic flux set up by circuit 1 which cuts circuit 2 (or conversely) and if N_1 and N_2 are the number of turns of coils 1 and 2, respectively, the interlinking flux is given by either of the relations

$$\phi_{21} = KN_2I_1 \qquad \text{or} \qquad \phi_{12} = KN_1I_2$$

The coefficient of mutual inductance, M, is defined as the number of flux linkages of one circuit per unit current flowing in the other. Hence, mathematically, the coefficient of mutual inductance is given by

$$M = \frac{N_1\phi_{21}}{I_1} = \frac{N_2\phi_{12}}{I_2} = KN_1N_2$$

The magnitude of the mutual inductance between two circuits depends upon the number of turns in each of the two circuits as well as upon the shapes of the coils, their separation (linear as well as angular), and other factors having a bearing on the flux linkages.

Mutual inductance may be derived from the basic field concepts. The calculations are long and not particularly easy to carry out for any particular physical configuration of conductors. A treatment of this subject may be found in Harnwell.[1]

The units in which mutual inductance is measured are the same as those for self-inductance: the henry and its decimal submultiples. But, whereas self-inductance is a positive quantity, mutual inductance may be either positive or negative.

140. Power Dissipated in a Resistor. If a potential difference, E, is maintained across a semiconductor (resistor), electric charges in the resistor will be set in motion. If a charge, dQ, is set in motion by the potential difference, E, the energy expended in the process is $dU = E\,dQ$, and if this element of charge traverses the resistor in time dt, the power expended in the process is

$$P = \frac{dU}{dt} = E\frac{dQ}{dt} = EI$$

For linear resistors that obey Ohm's law, $E = IR$; hence, the power dissipated in such a resistor is given by

$$P = EI = I^2R = \frac{E^2}{R}$$

The power dissipated will be measured in watts if I is the current in amperes, E is the potential drop in volts, and R is the resistance in ohms.

141. Energy Stored in Dielectric of a Capacitor. Suppose that at any instant of time the potential difference between the conducting elements of a capacitor has built up to the value E as a result of transferring an element of charge, dQ, to the plates or conductors. The amount of energy required for the process is $dU = E\,dQ$. Since, by definition of capacitance, $C = Q/E$, it follows that $Q = CE$ or $dQ = C\,dE$. Hence the energy increment is $dU = E\,dQ = CE\,dE$. Upon integrating, the total energy is found to be

$$U = \int dU = \int_0^E CE\,dE = 0.5CE^2 = 0.5\frac{Q^2}{C}$$

The energy, in mks units, is in joules, E is in volts, Q in coulombs, and C in farads.

142. Energy Stored in Magnetic Field of Inductor. From the relations given in Sec. 137, the drop in potential difference across an inductor is given by the relation $E = L(dI/dt)$, from which is obtained the expression $E\,dt = L\,dI$. The energy expended in time dt is given by the relation $dU = EI\,dt = LI\,dI$. Hence, the total energy in the magnetic field of an inductor is the integral of dU, or

$$U = \int dU = \int_0^I LI\,dI = 0.5LI^2 = 0.5L\left(\frac{dQ}{dt}\right)^2$$

In mks units, the energy is in joules, I is in amperes, L is in henrys, and Q is in coulombs.

[1] Harnwell, G. P., "Principles of Electricity and Electromagnetism," p. 297, McGraw-Hill, 1938.

The voltage-current relations for linear circuit elements such as have been treated here are summarized in Table 11, Sec. 155.

143. Circuit Laws and Theorems. The behavior of electrical circuits can be predicted on the basis of two relations originally stated by Kirchhoff and applicable under all conditions and a relation, originally stated by Ohm, which, although not universally applicable (at least in its simplest form), is true for many instances and is highly important and useful. In addition to these Kirchhoff and Ohm relations, a number of theorems find extensive usefulness in simplifying the calculations of circuit behavior. The most important of these are the principle of superposition, the compensation (substitution) theorem, the reciprocity theorem, the maximum power theorem, Thévenin's theorem, and Norton's theorem. These laws and theorems are treated in the following sections.

144. Kirchhoff's Laws. The two circuit laws of Kirchhoff may be stated as follows:

1. The algebraic sum of the currents flowing toward any point in an electric circuit is zero.
2. The algebraic sum of the product of the current and resistance in each of the conductors in any closed path in an electric circuit is equal to the algebraic sum of the emfs in that path.

As stated here, Kirchhoff's laws apply to the instantaneous values of the current and potential differences, but they may be extended to the effective values of sinusoidal currents and voltages by replacing "algebraic sum" by "vector sum" and replacing the term "resistance" by the term "impedance."

Each of the two Kirchhoff laws gives rise to a separate method of analyzing circuit behavior: the node and the loop method. In the node or junction method, the flow of current of all conductors meeting at a common point must be zero. Hence, if the magnitude and direction of all currents but one are known (or can be assumed), the remaining current can be determined by the application of the Kirchhoff law for current at a junction. In the loop method, the sum of all potential differences must be zero, or the sum of the rises of potential difference must equal the sum of the drops in potential difference.

145. Ohm's Law. Ohm's law states that the current in an electric circuit is directly proportional to the emf in the circuit. This law does not apply to all circuits but to metallic circuits and to many circuits containing electrolytic resistance. Ohm's law was first enunciated for a circuit in which there is constant emf and an unvarying current. It is applicable to circuits in which varying currents flow if account is taken of the induced emf resulting from the self-inductance of the circuit and of the distribution of current in the cross section of the circuit. Ohm's law applies only to linear circuits.

146. Principle of Linear Superposition. According to the principle of superposition, if several causes act simultaneously in a linear system, the resultant effect of all of them is equal to the sum of the individual effects of each cause acting by itself. As applied to electric circuits, the principle may be stated more explicitly. In a linear circuit, each emf acts independently of all other emfs in producing currents in the circuit.

147. Compensation Theorem. The compensation (or substitution) theorem states the conditions under which an emf may be replaced, in an electric circuit, by a potential difference across a resistor, so far as the circuit currents are concerned. Thus, provided the emf equals the potential drop across the resistance and is oppositely directed, a resistance element in a circuit can be replaced by an emf, or vice versa.

148. Reciprocity Theorem. In any branch, A, of an electric circuit, the current produced by an emf acting in another branch, B, is equal to the current in branch B when its emf is transferred to branch A. This is the equivalent of saying that a resistanceless source of emf may be interchanged with a resistanceless current-measuring instrument without altering the reading or indication of the instrument.

149. Maximum Power Transfer Theorem. The maximum amount of electric power is delivered from a source to a load connected to it when the resistance of the load is equal to the internal resistance of the power source. If the source and load contain

reactance as well as resistance, maximum power is transferred from the source to the load when the load resistance is equal to the internal resistance of the source and when the reactance of the load is equal in magnitude but opposite in sign to that of the source. Thus, maximum power is delivered from a source to a load when the load impedance is the conjugate of that of the source.

150. Thévenin's Theorem. If two terminals are available for any linear network, then, so far as external calculations of the system are concerned, they may be regarded as the terminals of a simple series circuit having a resistance equal to the resistance measured between the terminals, and an emf equal to the difference of potential between the two terminals.

151. Norton's Theorem. In a linear circuit, any generator of electric power may be considered equivalent, at specified frequency, to a current generator whose current is equal to the short-circuit current in shunt with an admittance whose magnitude is equal to that which is measured across the generator when the generator is inactive and there is no load connected to it.

152. Electric Networks. An electric network is a system of interconnected circuit elements and sources of emf. The points at which two or more circuit elements are connected are called "nodes" or "junctions." A continuous path (or several conducting paths) through which current may flow is called a "loop."

153. Circuit Connections. Fundamentally, two types of circuits are possible, according to the way in which the circuit elements are joined:

1. In series-connected circuits, the circuit elements are connected, one after the other, in such a way that the same current flows, successively, through each element, and the applied emf is equal to the sum of the drops in potential differences across all the successively connected circuit elements.

2. In parallel- or shunt-connected circuits the circuit elements are connected side by side in such a way that the same emf is applied simultaneously to each and every element, the total current flowing through the circuit being the sum of the currents flowing through each circuit element individually.

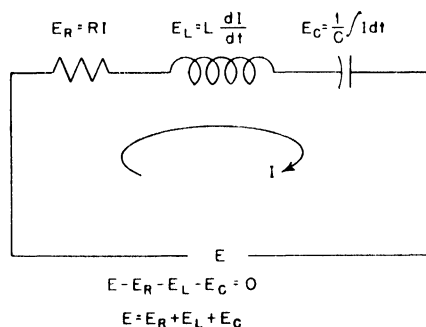



FIG. 32. Series circuit containing R, L, and C.

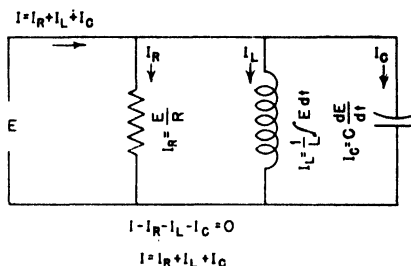



FIG. 33. Parallel circuit containing R, L, and C.

Examples of the most general two-terminal series-connected circuit and the most general two-terminal shunt-connected circuit containing, in each case, one of the three common circuit elements are shown in Figs. 32 and 33, respectively.

Combinations of series-connected circuit elements and shunt-connected circuit elements are possible to give rise to series-shunt or shunt-series circuits of any degree of complexity, but in any case only these two methods of connection are possible for two-terminal circuit elements.

154. Resultant of Series- and Shunt-connected Circuit Elements. Circuit elements connected either in series or in parallel, provided they are all of the same kind, may be replaced by an equivalent circuit element whose circuit parameter is the resultant

of that of the individually connected circuit elements. So long as there is no interaction between the fields of two or more circuit elements, *i.e.*, so long as no coupling exists between two or more circuit elements, the resultant circuit parameter of several like circuit elements connected together is simply expressed.

In the case of series-connected circuit elements, the resultant of several like circuit elements may be determined from the loop equations of Kirchhoff's laws. If subscripts 1, 2, . . . , n designate the component circuit elements connected in series without coupling or field interaction, the resultant circuit parameter, indicated by the subscript T, is given by the following relations:

For resistors,

$$R_T = R_1 + R_2 + R_3 + \cdots + R_n = \sum_{n=1}^{n} R_n$$

For inductors having no mutual coupling,

$$L_T = L_1 + L_2 + L_3 + \cdots + L_n = \sum_{n=1}^{n} L_n$$

For elastors (reciprocal capacitors) having no coupling,

$$S_T = S_1 + S_2 + S_3 + \cdots + S_n = \sum_{n=1}^{n} S_n$$

In the case of parallel-connected circuit elements, the resultant of several like circuit elements may be determined from the node or junction equation of Kirchhoff's laws. If the subscript notation is the same as that used above for series circuits, the resultant circuit parameter, indicated by the subscript T, is given by the following relations:

For conductance elements (reciprocal resistors),

$$G_T = G_1 + G_2 + G_3 + \cdots + G_n = \sum_{n=1}^{n} G_n$$

For reciprocal inductors having no mutual coupling,

$$\Gamma_T = \Gamma_1 + \Gamma_2 + \Gamma_3 + \cdots + \Gamma_n = \sum_{n=1}^{n} \Gamma_n$$

For capacitors having no coupling,

$$C_T = C_1 + C_2 + C_3 + \cdots + C_n = \sum_{n=1}^{n} C_n$$

In each case the resultant circuit element has properties like those of its constituent equivalent elements and is measured in the same electrical units of ohms, henrys, and darafs for the series circuits, and in mhos, yrnehs, and farads for the shunt-connected circuit elements, respectively.

155. Circuit Solutions. The solution for an electric circuit consists in finding the relations between the voltages (rises as well as drops of potential difference) and currents as a function of time. This relationship is determined by means of Kirchhoff's laws and by writing the node or loop equations for the circuit in terms of the general voltage-current relations for circuit elements as given in Table 11, and including any sources of emf that may exist in the circuit.

Since the general voltage-current relations for circuits composed of linear circuit elements are expressible through the use of differential, algebraic, or integral expressions, it is possible to differentiate these expressions with respect to time so that the

circuit behavior is expressed in terms of differential equations that represent the dynamic equilibrium of the circuit.

The complete solution of the differential equation representing the dynamic equilibrium of the electric circuit consists of two parts: (1) a transient (represented or expressed by the particular integral solution of the differential equation) specifying the temporary redistribution of charges (hence, a temporary voltage and current component) resulting from a sudden change in dynamic conditions, and (2) a steady-state solution (expressed by the complementary function of the differential equation) which represents the dynamic equilibrium of the circuit when subjected to an applied emf.

The currents and voltages representing the transient conditions of a network usually decrease rapidly with time and soon become negligible. In the steady-state condition, the voltages and currents may vary cyclically, but their effective value is constant.

Table 11. Voltage-Current and Energy Relations for Circuit Elements

Relation	Relation for the circuit property of		
	Resistance	Inductance	Capacitance
Voltage-current	$E = RI$	$E = L\frac{dI}{dt}$	$E = \frac{1}{C}\int I\,dt$
Current-voltage	$I = \frac{E}{R}$	$I = \frac{1}{L}\int E\,dt$	$I = C\frac{dE}{dt}$
Power dissipation	$P = I^2R = \frac{E^2}{R}$	$P = 0$	$P = 0$
Energy storage	$U = 0$	$U = \frac{1}{2}LI^2$	$U = \frac{1}{2}CE^2$

In the case of a harmonically applied emf, the drops in potential difference across each of the circuit elements and the current through the linear circuit elements are also harmonic functions of time and of the same frequency (but not the same phase, in general) as that of the applied emf. It is not generally true for any wave form, however, that the same wave shape exists for each of the circuit elements in the circuit.

156. Steady-state Solution. The steady-state condition of an electric circuit composed of linear two-terminal elements of R, L, and C is that in which the effective value of the current flowing in the circuit as a result of an applied emf is constant, each cycle of the recurrent wave form being like that of every other cycle. In general, the wave forms of the voltages across the different kinds of circuit elements of a series-connected circuit are unlike those across other circuit elements and, except in the case of a resistor, are not like the wave form of the applied emf. Likewise in shunt-connected circuits, the wave form of current through each circuit element is, in general, different from that flowing through circuit elements of another kind. Except for phase displacements, the wave forms are alike in the special but important case of harmonic voltages and currents. This is a direct consequence of the fact that the derivative and integral of a harmonic function are also harmonic functions.

The determination of the steady-state solution of electric circuits composed of linear circuit elements will be illustrated by finding the current that flows in a series R, L, C circuit when a harmonic emf, $E = E_m e^{j\omega t}$, is impressed across the series circuit of Fig. 32. From Kirchhoff's voltage or loop law and the voltage-current relations for the three kinds of circuit elements in the system, it follows that the loop equation takes the form

$$E = E_m e^{j\omega t} = E_L + E_R + E_C = L\frac{dI}{dt} + RI + \frac{1}{C}\int I\,dt$$

where I is the current flowing in the circuit and whose value is to be determined. Since it is easier to manipulate differential equations than integro-differential equa-

tions, the loop equation above may be transformed into a differential equation by differentiating both sides with respect to t. The result is

$$j\omega E_m e^{j\omega t} = L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}I$$

A solution of this equation is $I = I_m e^{j\omega t}$, from which $(dI/dt) = j\omega I_m e^{j\omega t}$ and $(d^2I/dt^2) = j^2\omega^2 I_m e^{j\omega t}$. Substituting these values in the above equation and dividing both sides by $j\omega$, there is obtained

$$E_m e^{j\omega t} = \left(j\omega L + R + \frac{1}{j\omega C}\right) I_m e^{j\omega t} = ZI_m e^{j\omega t}$$

where $Z = [j\omega L + R + (1/j\omega C)] = R + j[\omega L - (1/\omega C)]$ is the impedance of the circuit. From this result it is evident that current and voltage both vary harmonically with time, as does the voltage drop across each of the circuit elements.

157. Transient Solution. The transient condition of an electric circuit composed of linear elements of R, L, and C is that which involves the temporary redistribution of charges as the result of a sudden alteration of the dynamic equilibrium of the electric circuit. Since this transient condition expresses a readjustment of the dynamic equilibrium circuit, the solution does not involve any continuously externally applied emf or current. It does, however, involve the initial conditions of the circuit at the time when the circuit equilibrium is disturbed. The wave forms of voltage and current depend not only upon what elements are present in the circuit but also upon the relative magnitudes of the parameters of the circuit elements in the system.

The transient solution of electric circuits will be illustrated by finding the transient current that flows in a series circuit of linear R, L, and C circuit elements, when the quiescent equilibrium of the system is suddenly disturbed by closing the circuit with a battery of time-invariant emf, E, in the circuit. Physically, the circuit represents the charge on a capacitor in a circuit also containing inductance and resistance, under the assumption that the initial charge on the capacitor is zero.

At any instant of time, the Kirchhoff loop equation requires that

$$E = E_L + E_R + E_C = L\frac{dI}{dt} + RI + \frac{1}{C}\int I\,dt$$

To transform this to a differential equation, differentiate with respect to t to obtain

$$\frac{dE}{dt} = 0 = L\frac{d^2I}{dt^2} + R\frac{dI}{dt} + \frac{1}{C}I$$

the general solution of which is

$$I = \left[\frac{E}{\left(R^2 - \frac{4L}{C}\right)^{1/2}}\right] \epsilon^{-(Rt/2L)} \left[\epsilon^{\frac{1}{2L}\sqrt{R^2 - \frac{4L}{C}}\,t} - \epsilon^{-\frac{1}{2L}\sqrt{R^2 - \frac{4L}{C}}\,t}\right]$$

From this solution, it is evident that the term with the negative exponent indicates that the transient current will approach zero for a sufficiently large value of time, approaching zero current at infinite time. But the manner in which this zero current is approached depends upon the relative values of R, L, and C, and three different cases may exist for this circuit:

1. If $R^2 > 4L/C$, the exponential terms in the brackets are positive numbers; the current begins at zero, rises to a maximum value, and then gradually decays to zero slowly, without changing polarity.

2. If $R^2 = 4L/C$, the solution above becomes indeterminate, but by differentiating both numerator and denominator individually with respect to the factor that is the coeffi-

cient of t, the solution for the current is

$$I = \left(\frac{Et}{L}\right) e^{-(Rt/2L)}$$

For this case, often called the "critical" case, the current begins at zero, rises to a maximum value, and returns again to zero without changing polarity. This is the limiting condition of case 1 in which the rise and fall occur more slowly than for the present critical case.

3. If $R^2 < 4L/C$, the current takes the form

$$I = \left(\frac{2E}{\sqrt{\frac{4L}{C} - R^2}}\right) e^{-(Rt/2L)} \sin \omega t$$

where $\omega = \frac{1}{2L}\left(\frac{4L}{c} - R^2\right)^{\frac{1}{2}}$. For this case it is evident from the equation above that the current oscillates with an angular velocity ω and that the amplitude steadily decreases with time until the magnitude of the transient oscillations becomes negligible, ultimately reaching zero at infinite time.

Transients are assuming increasing importance in radio engineering design, and the following list of references will serve as a suitable introduction to this topic.

References

Berg, E. J.: "Heaviside's Operational Calculus," McGraw-Hill, 1936.
Bush, V.: "Operational Circuit Analysis," Wiley, 1929.
Carson, J. R.: "Electric Circuit Theory and the Operational Calculus," McGraw-Hill, 1926.
—— —— and O. J. Zobel: Transient Oscillations in Electric Wave Filters, *Bell System Tech. J.*, **2** (3), 1, July, 1923.
Gardner, M. F., and J. L. Barnes: "Transients in Linear Systems," Wiley, 1942.
Guillemin, E. A.: "Communication Networks," Vol. 1, Wiley, 1931.
Kerchner, R. M., and G. F. Corcoran: "Alternating Current Circuits," Wiley, 1938.
Kurtz, E. B., and G. F. Corcoran: "Introduction to Electric Transients," Wiley, 1935.
Massachusetts Institute of Technology, Electrical Engineering Dept.: "Electric Circuits," Wiley, 1940.
McIlwain, K., and J. G. Brainerd: "High Frequency Alternating Currents," Wiley, 1931.
Pierce, G. W.: "Electric Oscillations and Electric Waves," McGraw-Hill, 1920.
Skilling, H. H.: "Transient Electric Currents," McGraw-Hill, 1937.

158. Impedance, Resistance, and Reactance. The impedance of an electric circuit, or portion of an electric circuit, to a completely specified periodic potential difference and current is the ratio of the effective or rms value of the potential difference across the circuit to the effective value of the current flowing through it, there being no source of power in the portion under consideration. The impedance is designated by the letter symbol Z, and if E_{rms} and I_{rms} are the effective values of the potential drop and the current of the circuit, the circuit impedance is given by

$$Z = \frac{E_{\text{rms}}}{I_{\text{rms}}}$$

The impedance is measured in ohms.

If the periodic emf applied to the circuit is composed of components of different frequency, if $E_0, E_1, E_2, \ldots, E_n$ represent the effective values of voltages of frequencies $0, \omega, 2\omega, \ldots, n\omega$, if $I_0, I_1, I_2, \ldots, I_n$ represent the effective values of the currents of the same frequencies, respectively, and if the maximum values of the voltage are $E_0, E_{1m}, E_{2m}, \ldots, E_{nm}$ and the maximum (amplitude) values of the currents are $I_0, I_{1m}, I_{2m}, \ldots, I_{nm}$, then the impedance of the circuit or portion of the circuit is given by the relation

$$Z = \frac{E_{\text{rms}}}{I_{\text{rms}}} = \frac{\sqrt{2E_0^2 + E_{1m}^2 + E_{2m}^2 + \cdots}}{\sqrt{2I_0^2 + I_{1m}^2 + I_{2m}^2 + \cdots}} = \frac{\sqrt{E_0^2 + E_1^2 + E_2^2 + \cdots}}{\sqrt{I_0^2 + I_1^2 + I_2^2 + \cdots}}$$

The impedance thus defined is a scalar quantity and gives merely the ratio of the magnitudes of the voltage and current.

The vector impedance of an electric circuit, or portion of a circuit, for simple harmonic voltage and current is the ratio of the corresponding complex harmonic potential difference to the corresponding complex current. In the example given in Sec. 156, the applied harmonic voltage is $E = E_m e^{j\omega t}$ and the resulting current is $I = I_m e^{j\omega t}$ and the vector impedance, *i.e.*, the impedance expressed as a vector in rectangular coordinates, is

$$Z = \frac{E_m e^{j\omega t}}{I_m e^{j\omega t}} = \left[R + \left(j\omega L + \frac{1}{j\omega C}\right)\right] = \left[R + j\left(\omega L - \frac{1}{\omega C}\right)\right] = R \pm jX$$

where R is the resistance, $j\omega L$ is the inductive reactance in ohms, $1/j\omega C = -j/\omega C$ is the capacitive reactance in ohms, and $X = j\left(\omega L - \frac{1}{\omega C}\right)$ is the reactance in ohms. Depending upon whether the inductive or the capacitive reactance has the larger magnitude, the net reactance will be, respectively, positive or negative, but the resistance will always be positive for linear dissipative circuits.

In the above expression the impedance has been given in rectangular components, R and X, but the impedance may also be expressed in polar coordinates in terms of the magnitude of the impedance, $|Z|$, and the angle of phase shift, $\underline{/\theta}$, of the impedance. Thus, in polar coordinates, the impedance for the series R, L, C circuit is

$$Z = |Z|\underline{/\theta} = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \Bigg/ \tan^{-1} \frac{R}{\left(\omega L - \frac{1}{\omega C}\right)}$$

$$= \sqrt{R^2 + X^2}\ \underline{/\tan^{-1}}\ (R/X)$$

where $|Z| = \sqrt{R^2 + X^2}$ and $\theta = \tan^{-1}(R/X)$. Conversely, the rectangular components of the impedance, in terms of the polar components, are

$$R = |Z| \cos \theta$$
$$X = |Z| \sin \theta$$

Physically the impedance of a circuit is the opposition to the flow of a periodic current under steady-state conditions. It is composed of two parts: a resistive component usually regarded as constant (although the resistance is, strictly speaking, a function of frequency) and a reactive component whose magnitude usually varies appreciably with frequency. Over a sufficiently large frequency range the sign of the reactance also changes. The resistive component accounts for the absorption or dissipation of power. The reactive component results from properties of inductors and capacitors that involve the energy storage and rate-of-change relationships between voltage and current. For ideal circuit elements the reactance determines the frequency-selective properties of the circuit.

The impedance is an electrical quantity whose use is most convenient in solving electrical circuits with circuit elements connected in series.

159. Admittance, Conductance, and Susceptance. The admittance of an electric circuit is the reciprocal of the impedance; it is, therefore, the ratio of the effective value of the current through a circuit or portion of a circuit to the effective value of the potential difference across the circuit, for a completely specified periodic applied voltage. The admittance is designated by the letter Y and is measured in reciprocal ohms or mhos or its multiples or submultiples.

If E_{rms} and I_{rms} are the effective values of the voltage and current, respectively, the admittance is given by

$$Y = \frac{I_{\text{rms}}}{E_{\text{rms}}}$$

For a nonharmonic periodic voltage, the admittance may be determined in a manner similar to that for determining the impedance of a circuit in which nonharmonic current flows.

In any electric circuit whose resistive and reactive components are R and X, respectively, the corresponding admittance is given by the relation

$$Y = \frac{1}{Z} = \frac{1}{R \pm jx} = \frac{1}{R \pm jx} \cdot \frac{R \mp jx}{R \mp jx} = \frac{R}{R^2 + X^2} \mp j \frac{X}{R^2 + X^2} = G \mp jB$$

where $G = R/(R^2 + X^2) = R/Z^2$ is the conductance of the circuit (measured in mhos) and $B = \pm X/(R^2 + X^2) = X/Z^2$ is the susceptance of the circuit (measured in mhos).

The admittance may be expressed in polar components as well as in the rectangular components given above. In polar components,

$$Y = |Y|\underline{/\phi} = \sqrt{G^2 + B^2}\ \underline{/\tan^{-1} (B/G)}$$

where $|Y| = \sqrt{G^2 + B^2}$ and $\phi = \tan^{-1} (B/G)$. Conversely, the rectangular components of the admittance, in terms of polar components, are

$$G = |Y| \cos \phi$$
$$B = |Y| \sin \phi$$

Since impedance and admittance are reciprocally related, the impedance may be expressed in terms of admittance components just as admittance was expressed, above, in terms of impedance components. Thus, in terms of the admittance, the impedance is

$$Z = \frac{1}{Y} = \frac{1}{G \pm jB} = \frac{1}{G \pm jB} \cdot \frac{G \mp jB}{G \mp jB} = \frac{G}{G^2 + B^2} \mp j \frac{B}{G^2 + B^2}$$

Physically, the admittance expresses the ease with which the current flows through a circuit. It is composed of two parts: a conductance component, which is usually regarded as constant and which enters into determinations of power, and a susceptive component, which usually varies appreciably with frequency in a manner inverse to that of the reactance of the circuit.

Admittance is a concept that is most useful in dealing with parallel circuits.

160. Resonance. Resonance exists between one coordinate of a system that is executing oscillations or vibrations and a periodic agency that maintains the oscillations or vibrations when a small amplitude of the periodic agency produces in the system a relatively large amplitude of the oscillation or vibration.

Amplitude resonance exists when the resonance is such that any change in the period of the periodic agency without changing its amplitude produces a decrease in the amplitude of the oscillation or vibration of the system.

Period or natural resonance exists when the resonance is such that the period of the applied agency is the same as the natural period of oscillation or vibration of the system.

Phase or velocity resonance exists when the resonance is such that the angular phase difference between the fundamental components of the oscillation or vibration and of the applied agency is 90 deg ($\pi/2$ radians). With phase resonance, the time derivative of the coordinate that is in resonance is a maximum.

Resonance may exist in an electric circuit possessing inductance, resistance, and capacitance between the quantity of electricity that oscillates and a periodically applied emf that sustains the oscillations. Any one of three kinds of resonance, *viz.*, amplitude resonance, period resonance, or phase resonance, may exist depending upon the constants of the circuit. In electric circuits having a high degree of selectivity, there is little distinction between the three kinds of resonance.

In a series circuit composed of linear circuit elements of R, L, and C, resonance is obtained when the net reactance is 0 or when $\omega L = 1/\omega C$. Hence, the frequency of resonance is obtained from the relation $\omega^2 = 1/(LC)$ so that

$$f_r = \frac{1}{2\pi \sqrt{LC}}$$

The frequency will be expressed in cycles per second if L is inductance in henrys and C is capacitance in farads.

References

Everitt, W. L., and A. E. Anner: "Communication Engineering," McGraw-Hill, 1956.
Gardner, M. F., and J. L. Barnes: "Transients in Linear Systems," Wiley, 1956.
Glasgow, R. S.: "Principles of Radio Engineering," McGraw-Hill, 1936.
Guillemin, E. A.: "Communication Networks," Vols. I and II, Wiley, 1931 and 1935.
———: "Mathematics of Circuit Analysis," Wiley, 1949.
———: "Synthesis of Passive Networks," Wiley, 1957.
Hague, B.: "Alternating Current Bridge Methods," Pitman, 1943.
Johnson, K. S.: "Transmission Circuits for Telephone Communication," Van Nostrand, 1925.
Kerchner, R. M., and G. F. Corcoran: "Alternating Current Circuits," Wiley, 1943.
Knight, A. R., and G. H. Fett: "Introduction to Circuit Analysis," Harper, 1943.
Martin, T. L.: "Physical Basis for Electrical Engineering," Prentice-Hall, 1957.
Massachusetts Institute of Technology, Electrical Engineering Dept.: "Electric Circuits," Wiley, 1940.
McIlwain, K., and J. G. Brainerd: "High Frequency Alternating Currents," Wiley, 1931.
Shea, T. E.: "Transmission Networks and Wave Filters," Van Nostrand, 1929.
Starr, A. T.: "Electric Circuits and Wave Filters," Pitman, 1934.
Stewart, J. L.: "Circuit Theory and Design," Wiley, 1950.
Terman, F. E.: "Electronic and Radio Engineering," McGraw-Hill, 1955.
Ware, L. A., and M. R. Reed: "Communication Circuits," Wiley, 1944.

MAGNETIC CIRCUITS

161. Electric and Magnetic Circuits. A magnetic circuit is a closed path of magnetic flux, the path having the direction of the magnetic induction at every point. On the other hand, an electric circuit is a closed path of electric current, the path having the direction of the electric current at every point. Although the equations for electric and magnetic circuits may take the same mathematical form, the analogy between them expresses no physical correspondence. Particularly is this true for magnetic induction, which does not represent a dissipation of power, whereas the flow of electric current does.

Corresponding to the relation between voltage, current, and resistance, as given by Ohm's law for an electric circuit, the magnetic circuit relation may be written

$$\mathfrak{F} = \phi \mathfrak{R}$$

where $\mathfrak{F}$ is the magnetomotive force in ampere-turns in the mks system, ϕ is the magnetic flux or induction in webers, and $\mathfrak{R}$ is the reluctance of the magnetic circuit in ampere-turns per weber. An unvarying magnetic flux is regarded as a static condition, whereas in the equivalent electric circuit, the steady current is the uniform flow of electric charges. Reluctance is not analogous to resistance in the sense of accounting for energy dissipation. Another difference between electric and magnetic circuits is the degree to which electric or magnetic effects are confined to their respective circuits. Whereas excellent insulators for the flow of electric conduction current exist, there is no insulator for magnetic induction; therefore, the magnetic flux is not so completely confined to its magnetic circuit as the conduction current is confined to the electric circuit. From the practical point of view, the most useful magnetic circuits consist mainly of ferromagnetic materials, and for such materials, the reluctance is not constant; in this respect too, magnetic circuits differ from most electric circuits for which Ohm's law is obeyed.

162. Magnetic Permeability and Reluctivity. Permeability is the property of an isotropic medium which, under specified conditions, determines the relation between the magnitudes of the magnetizing force and the magnetic induction of the medium. Under the specified conditions, permeability is measured as the ratio of the magnetic induction to the magnetizing force, or

$$\mu = \frac{\mathfrak{B}}{\mathfrak{H}}$$

where, in mks units, $\mathfrak{B}$ is the magnetic flux density or induction in webers per square meter, and $\mathfrak{H}$ is the magnetic field intensity in ampere-turns per meter.

The relative permeability of most substances does not differ appreciably from that of vacuum, which is taken as unity. For practical purposes the relative permeability is constant, independent of the flux density. Iron, nickel, steel, cobalt, and magnetic alloys—the materials of practical importance in magnetic circuits—have relative permeabilities appreciably greater than unity, and the permeability varies considerably with the magnetizing force or the flux density, with the composition, and with the previous heat-treatment of the magnetic material.

For a uniform magnetic path, the reluctance may be determined from the field expressions for the magnetomotive force and magnetic flux,

$$\mathcal{R} = \frac{\mathcal{F}}{\phi} = \frac{\int \mathcal{H} \cos\theta \, dl}{\iint_s \mathcal{B} \cos\theta' \, ds} = \frac{\mathcal{H} \int dl}{\mu \mathcal{H} \iint_s ds} = \frac{l}{\mu s} = \nu \frac{l}{s}$$

where $\mathcal{F}$ is the magnetomotive force in ampere-turns, ϕ is the magnetic flux in webers, $\mathcal{H}$ is the magnetic field intensity in ampere-turns per weber, θ is the angle between $\mathcal{H}$ and l or the angle between the direction of $\mathcal{B}$ and the normal to the surface s and for this case has the value $\theta = 0$, μ is the permeability in henrys per meter, and $\nu = 1/\mu$ is the reluctivity in meters per henry. For nonmagnetic substances (vacuum, air, and so on) the permeability and reluctivity are usually written μ_v and ν_v.

In some magnetic circuits, there are varying as well as steady components of magnetic flux. The permeability offered to the varying component of the magnetizing force superimposed upon the steady component of the magnetizing force is called the "incremental permeability," the "apparent permeability," or sometimes the "a-c permeability." For such cases the magnetization cycle follows a small displaced hysteresis loop. The incremental permeability is measured by the slope of the straight line connecting the end points of this hysteresis loop. The incremental permeability decreases with increasing magnetization due to the steady component of magnetizing force and, up to the point at which saturation occurs, increases with the magnitude of the time-varying component of the magnetic flux density.

163. Reluctance and Permeance in Series and Parallel Magnetic Circuits. Permeance of a magnetic circuit is the reciprocal of the reluctance and expresses the ease with which magnetic flux may be established in a magnetic circuit, whereas reluctance is a measure of the opposition or difficulty with which magnetic flux is established in the magnetic circuit.

For magnetic circuits in series with one another, the total reluctance is the sum of the reluctances of the individual circuits, plus the reluctance of the joints where the magnetic circuits of high permeability are joined. If magnetic circuits are placed in parallel with one another, the permeance of each circuit is added to ascertain the total or net permeance of the combination of magnetic circuit branches in parallel.

References

Abraham, M., and R. Becker: "Classical Theory of Electricity and Magnetism," Blackie, 1932.
Bitter, F.: "Ferromagnetism," McGraw-Hill, 1937.
Ewing, J. A.: "Magnetic Induction in Iron and Other Substances," Van Nostrand, 1900.
Massachusetts Institute of Technology, Electrical Engineering Dept.: "Magnetic Circuits and Transformers," Wiley, 1943.
Moullin, E. B.: "Principles of Electromagnetism," Oxford, 1932.
Spooner, T.: "Properties and Testing of Magnetic Materials," McGraw-Hill, 1931.
Stoner, E. C.: "Magnetism and Matter," London, 1934.
Williams, S. R.: "Magnetic Phenomena," McGraw-Hill, 1931.

ELECTROMAGNETIC WAVES

164. Electromagnetic Waves. A wave is a disturbance that is propagated in a medium in such a way that, at any point in the medium, the displacement of the wave is a function of the position of the point. Any physical quantity that has the same relationship to some independent variable (usually time) that a propagated disturbance has at a particular instant, with respect to space, may be called a "wave." The acceleration of electric charges is capable of producing waves in the electromagnetic

field, and it is these electromagnetic waves that make possible radio communication without the intervention of physical conductors between the transmitting and the receiving stations.

Maxwell's equations express the general relations between electrical and magnetic quantities, *i.e.*, between $\mathcal{E}$, $\mathcal{D}$, $\mathcal{H}$, and $\mathcal{B}$; hence, the Maxwell equations may be expected to yield the means for deriving the equations of electromagnetic waves. This is indeed the case, but the variables are interconnected in such a manner that Maxwell's differential equations must be simplified (by separating the variables) before any solutions for electromagnetic waves can be ascertained. When the variables have been properly separated for the four Maxwell field equations through suitable mathematical manipulations, it is next required to find the relation or relations connecting time and the directional derivatives of $\mathcal{E}$, $\mathcal{D}$, $\mathcal{H}$, and $\mathcal{B}$, such that all four field equations are simultaneously satisfied and that the time-space relations will be in accordance with the definitions above for the propagation of electromagnetic waves. Since the first two (divergence) field equations do not involve time functions, the procedure is essentially that of finding a solution to the two circuital equations, both of which must be simultaneously satisfied.

In a medium that is perfectly insulating (zero conductivity) and free from charge density (free space satisfies these two conditions), the field equations to be solved, when time and space variables have been separated, are, in rectangular coordinates,

$$\mathbf{i}\frac{\partial \mathcal{E}_x}{\partial x} + \mathbf{j}\frac{\partial \mathcal{E}_y}{\partial y} + \mathbf{k}\frac{\partial \mathcal{E}_z}{\partial z} = 0$$

$$\mathbf{i}\frac{\partial \mathcal{H}_x}{\partial x} + \mathbf{j}\frac{\partial \mathcal{H}_y}{\partial y} + \mathbf{k}\frac{\partial \mathcal{H}_z}{\partial x} = 0$$

$$\mathbf{i}\left(\frac{\partial^2 \mathcal{E}_x}{\partial x^2} + \frac{\partial^2 \mathcal{E}_x}{\partial y^2} + \frac{\partial^2 \mathcal{E}_x}{\partial z^2}\right) = \mu\epsilon\frac{\partial^2 \mathcal{E}_x}{\partial t^2}$$

$$\mathbf{j}\left(\frac{\partial^2 \mathcal{E}_y}{\partial x^2} + \frac{\partial^2 \mathcal{E}_y}{\partial y^2} + \frac{\partial^2 \mathcal{E}_y}{\partial z^2}\right) = \mu\epsilon\frac{\partial^2 \mathcal{E}_y}{\partial t^2}$$

$$\mathbf{k}\left(\frac{\partial^2 \mathcal{E}_z}{\partial x^2} + \frac{\partial^2 \mathcal{E}_z}{\partial y^2} + \frac{\partial^2 \mathcal{E}_z}{\partial z^2}\right) = \mu\epsilon\frac{\partial^2 \mathcal{E}_z}{\partial t^2}$$

$$\mathbf{i}\left(\frac{\partial^2 \mathcal{H}_x}{\partial x^2} + \frac{\partial^2 \mathcal{H}_x}{\partial y^2} + \frac{\partial^2 \mathcal{H}_x}{\partial z^2}\right) = \mu\epsilon\frac{\partial^2 \mathcal{H}_x}{\partial t^2}$$

$$\mathbf{j}\left(\frac{\partial^2 \mathcal{H}_y}{\partial x^2} + \frac{\partial^2 \mathcal{H}_y}{\partial y^2} + \frac{\partial^2 \mathcal{H}_y}{\partial z^2}\right) = \mu\epsilon\frac{\partial^2 \mathcal{H}_y}{\partial t^2}$$

$$\mathbf{k}\left(\frac{\partial \mathcal{H}_z}{\partial x^2} + \frac{\partial^2 \mathcal{H}_z}{\partial y^2} + \frac{\partial^2 \mathcal{H}_z}{\partial z^2}\right) = \mu\epsilon\frac{\partial^2 \mathcal{H}_z}{\partial t^2}$$

Only electromagnetic waves that (when expressed in rectangular coordinates) satisfy these equations are possible; the equations above may, therefore, be used to determine the required form of the fields for wave propagation.

Solution of the above equations requires that each directional component of the electric or magnetic field intensity be of the form

$$\left.\begin{matrix}\mathcal{E}\\ \mathcal{H}\end{matrix}\right\} = f\left(t - \frac{l}{v}\right) + f\left(t + \frac{l}{v}\right)$$

where v is the velocity of propagation and is given by $v = 1/\sqrt{\mu\epsilon}$, l is used generically to designate a distance in any of the three orthogonal directions, and the functions, f, represent any single-valued functions. For propagation in free space, the propagation velocity has the special value of the velocity of light, approximately 3×10^8 m

per sec obtained from the relation $v = c = 1/\sqrt{\mu_0\epsilon_0}$. The fact that the functions, f, are arbitrary functions of time makes it possible for the electromagnetic waves to be formed in such a way as to be capable of representing intelligence The fundamental problem of radio communication is to make the time variations of the electromagnetic wave conform to the variations of the intelligence to be transmitted.

The energy stored in the electromagnetic field travels with the wave from the source of the disturbance outward into space, and this phenomenon is the mechanism by which the transmission of electric power, sufficient to carry on communication, occurs. An important point regarding the transmission of electromagnetic waves through space is that the energy is propagated through a dielectric; no conduction can occur in the (ideal) dielectric of free space. The mechanism of propagation is, briefly, that the variations of electric field intensity give rise to variations of magnetic field intensity which, in turn, give rise to variations of the electric field intensity once more, and so on. Once established, the original disturbance is sustained by the mutual interaction of electric and magnetic components of the electromagnetic field as the wave travels through space with the velocity of light. The amplitude of the variations of the electromagnetic field decreases with distance from the center of the disturbance as the wave expands to greater and greater size.

165. Energy in the Electromagnetic Field. Since, as noted above, the energy of the electromagnetic field radiates outward into space, it is desirable to have quantitative expressions for the field energy. Such expressions can be obtained from energy relations derived in Secs. 141 and 142 by converting the circuital relations into field relations.

From the relation giving the total energy stored in the dielectric of a capacitor, the total electric energy stored in an electromagnetic field of volume v is

$$U_{\mathcal{E}} = 0.5CE^2 = 0.5QE = 0.5K''\left(\iint_s \mathfrak{D}\cdot ds\right)\cdot\left(\int \mathcal{E}\cdot dl\right)$$
$$= 0.5K''\left(\iint_s \epsilon\mathcal{E}\cdot ds\right)\cdot\left(\int \mathcal{E}\cdot dl\right) = 0.5K''\epsilon\mathcal{E}^2\iiint dv = 0.5K''\epsilon\mathcal{E}^2 v$$

so that the electric energy per unit volume of field is

$$u_{\mathcal{E}} = 0.5K''\epsilon\mathcal{E}^2$$

In a similar way, by making use of the energy in the magnetic field surrounding a coil, the total magnetic energy stored in an electromagnetic field of volume v is

$$U_{\mathcal{H}} = 0.5LI^2 = 0.5\left(\frac{K''\iint_s \mathfrak{D}\cdot ds}{\int \mathcal{H}\cdot dl}\right)\cdot\left(\int \mathcal{H}\cdot dl\right)^2$$
$$= 0.5K''\left(\iint_s \mu\mathcal{H}\cdot ds\right)\cdot\left(\int \mathcal{H}\cdot dl\right) = 0.5K''\mu\mathcal{H}^2\iiint dv = 0.5K''\mu\mathcal{H}^2 v$$

so that the magnetic energy per unit volume is

$$u_{\mathcal{H}} = 0.5K''\mu\mathcal{H}^2$$

The total energy, per unit volume, is the sum of the electric and magnetic energy, or

$$u = u_{\mathcal{E}} + u_{\mathcal{H}} = K''(0.5\epsilon\mathcal{E}^2 + 0.5\mu\mathcal{H}^2)$$

166. Poynting's Vector. Poynting's vector has importance in the determination of power flow throughout space; it is, therefore, frequently encountered in radiation phenomena. Usually designated by the letter S, it represents the energy per second flowing per unit area perpendicular to the direction of propagation.

At a point in space, the total flow of power (or energy per unit time) is equal to the product of the electric field intensity and the current density (representing a dissipation of power) plus the time rates of change of the magnetic field and the electric field.

Mathematically, in terms of Poynting's vector, S,

$$-\text{div } S = \mathcal{J}\mathcal{E} + \frac{\partial u_{\mathcal{E}}}{\partial t} + \frac{\partial u_{\mathcal{H}}}{\partial t}$$

This equation defines conditions at a point in the electromagnetic field, but it is possible to define the same conditions for a volume, dv. Thus, by expressing $\mathcal{J}$, $\mathcal{E}$, and $\mathcal{H}$ in terms of their orthogonal components, we obtain the integral expression for rectangular coordinates,

$$-\iiint_v \left(\frac{\partial s_x}{\partial x} + \frac{\partial s_y}{\partial y} + \frac{\partial s_z}{\partial z}\right) dv = \iiint_v (\mathcal{J}_x\mathcal{E}_x + \mathcal{J}_y\mathcal{E}_y + \mathcal{J}_z\mathcal{E}_z)\, dv$$
$$+ \iiint_v \frac{\partial}{\partial t}\left(\frac{\epsilon\mathcal{E}_x^2}{2} + \frac{\epsilon\mathcal{E}_y^2}{2} + \frac{\epsilon\mathcal{E}_z^2}{2}\right) dv + \iiint_v \frac{\partial}{\partial t}\left(\frac{\mu\mathcal{H}_x^2}{2} + \frac{\mu\mathcal{H}_y^2}{2} + \frac{\mu\mathcal{H}_z^2}{2}\right) dv$$

167. Boundary Conditions for Electric and Magnetic Fields. In order that any electromagnetic field distribution may be realized physically, it must satisfy Maxwell's equations. In addition it must also satisfy boundary conditions for the physical system in which the electrical effects take place.

FIG. 34. Boundary conditions for electric and magnetic fields.

Whenever an electromagnetic field or wave meets a boundary between two mediums of different dielectric or magnetic properties, there is a change in the fields and, in general, the wave splits into two components, one of which is reflected back into the first medium, the other being refracted into the second medium. For any two mediums, 1 and 2, the tangential components of the electric field and also of the magnetic field are equal on both sides of the dividing surface. Hence, as shown in Fig. 34, the relation exists that

$$\mathcal{E}_{1_T} = \mathcal{E}_{2_T} \qquad \text{and} \qquad \mathcal{H}_{1_T} = \mathcal{H}_{2_T}$$

The relationship between the normal components may be determined from Gauss's law. If the surface charge density of the surface separating the two mediums is σ, then the relation exists that

$$\mathfrak{D}_{1_n} - \mathfrak{D}_{2_n} = \sigma$$

If, as is often the case, the surface charge density is zero, then the above relation simplifies to

$$\mathfrak{D}_{1_n} = \mathfrak{D}_{2_n}$$

From similar reasoning it can be shown that for the magnetic field, for which no free magnetic poles exist, the required relation is

$$\mathcal{B}_{1_n} = \mathcal{B}_{1_n}$$

References

Hund, A.: "Phenomena in High-frequency Systems," McGraw-Hill, 1936.
Jeans, J. H.: "Mathematical Theory of Electricity and Magnetism," Cambridge, 1943.
King, R. W. P.: "Electromagnetic Engineering," Vol. I, McGraw-Hill, 1945.
Maxwell, J. C.: "Treatise on Electricity and Magnetism," Oxford, 1904.
Morse, P. M., and H. Feshbach: "Methods of Theoretical Physics," McGraw-Hill, 1953.
Moullin, E. B.: "Principles of Electromagnetism," Oxford, 1932.

Pierce, G. W.: "Electric Oscillations and Electric Waves," McGraw-Hill, 1920.
Ramo, S., and J. R. Whinnery: "Fields and Waves in Modern Radio," Wiley, 1944.
Rawer, K.: "The Ionosphere," Fred'k Unger, 1957.
Schelkunoff, S. A.: "Electromagnetic Waves," Van Nostrand, 1943.
Skilling, H. H.: "Fundamentals of Electric Waves," Wiley, 1942.
Slater, J. C.: "Microwave Transmission," McGraw-Hill, 1942.
Stratton, J. A.: "Electromagnetic Theory," McGraw-Hill, 1941.

UNITS AND DIMENSIONS

168. Nature of Units. Quantities of many kinds (mechanical, thermal, chemical, physiological, etc.) are encountered in scientific and engineering work. The magnitudes of different quantities are recorded or compared with one another by selecting some magnitude as unity and expressing all others of the same quantity as a ratio to the unit quantity or magnitude. Hence, any quantity is composed of two parts: (1) a numeric which expresses the relative magnitude of the quantity in terms of the established unit magnitude, and (2) a statement designating the kind of measurement that the quantity represents. For example, the velocity of 3 m per sec contains these two parts: the numeric 3 indicates that the quantity in question has a magnitude 3 times that of unit magnitude in the system of measurement under discussion, and the kind of quantity is indicated by the measure of meters per second, having the dimensions of length divided by time.

Each quantity selected as a reference unit must be established by a physical standard of some kind, or, alternatively, it must be obtained or derived from combinations of other units that are selected or recognized as standard units. The units may be artificial or natural.

For engineering use, two general systems of mechanical units are in use in the United States. The English system of units, most frequently used, is based on the custom of common law and common and generally widespread use. The metric system of units, on the other hand, is legalized by an act of Congress of July 28, 1866. The metric system is extensively used in scientific work because of its simplicity and the ease with which decimally related quantities may be converted from one set of units to another.

169. English Units. Long usage and custom have lent a considerable weight of authority to the English system of units in the United States, in spite of the unnatural and unrational relationships that exist between the various quantities in this system. Because of the unnecessary complication of the oddly related units, the English system has found little support in scientific circles. In engineering work, which is more closely related to the everyday commercial transactions than is scientific work, the English system of units is encountered to such an extent that familiarity with it is requisite. The English system is based on the foot as the unit of length, the second as the unit of time, and the slug as the unit of mass.

170. Metric Units. The metric system, which finds extensive application throughout the scientific world, is based primarily upon the prototype standards of mass and length preserved at the International Bureau of Weights and Measures at Sèvres, France. Similar standards of length and mass are deposited in numerous national laboratories and serve to define the legal standards for all the nation.

Table 12. Metric Prefixes

Name of prefix	Symbol	Multiplying factor
mega-	M	1,000,000
kilo-	k	1,000
hecto-	..	100
deka-	..	10
......	..	1
deci-	..	0.1
centi-	c	0.01
milli-	m	0.001
micro-	μ	0.000,001

A great convenience of the metric system is that all units are expressed in decimal multiples or submultiples of the unit selected for the standard or reference value. The decimal multiples or submultiples are designated in Table 12. Combinations of these prefixes may be used to designate intermediate values; for example, millimicro may be used to designate 0.000,000,001.

Table 14. Electrical Units and Dimensions

Name	Symbol	Definition	Dimensions m^A A	l^B B	t^C C	Q^D D
Charge	Q	Fundamental $f = \frac{Q_1 Q_2}{\epsilon l^2}$	0	0	0	1
Surface charge density	σ	$\sigma = \lim_{\Delta s \to 0} \frac{\Delta Q}{\Delta s}$	0	−2	0	1
Volume charge density	ρ	$\rho = \lim_{\Delta V \to 0} \frac{\Delta Q}{\Delta V}$	0	−3	0	1
Electric field intensity	$\mathcal{E}$	$\mathcal{E} = \lim_{\Delta Q \to 0} \frac{\Delta f}{\Delta Q}$	1	1	−2	−1
Displacement	$\mathfrak{D}$	$\mathfrak{D} = \epsilon \mathcal{E}$	0	−2	0	1
Permittivity	ϵ	$\epsilon = \epsilon_0 \epsilon_r = \frac{\mathfrak{D}}{\mathcal{E}}$	−1	−3	2	2
Electric flux	ψ	$\psi = \int\int_s \mathfrak{D}_n \, ds$	0	0	0	1
Pole strength	M	$f = \frac{M_1 M_2}{\mu l^2}$	1	2	−1	−1
Magnetic field intensity	$\mathcal{H}$	$\mathcal{H} = \lim_{\Delta M \to \theta} \frac{\Delta f}{\Delta M} = (NIl \sin \theta)/r^2$	0	−1	−1	1
Magnetic flux density	$\mathcal{B}$	$\mathcal{B} = \mu \mathcal{H}$	1	0	−1	1
Magnetic flux	ϕ	$\phi = \int\int_s \mathcal{B}_n \, ds$	1	2	−1	−1
Permeability	μ	$\mu = \mu_0 \mu_r = \frac{\mathcal{B}}{\mathcal{H}}$	1	1	0	−2
Reluctivity	ν	$\nu = \frac{1}{\mu} = \frac{\mathcal{H}}{\mathcal{B}}$	−1	−1	0	2
Electric voltage (potential and induced electromotive force)	E, e	$E = \int \mathcal{E} \cdot dl + \frac{\partial}{\partial t} \int\int_s \mathcal{B}_n \, ds$	1	2	−2	−1
Capacitance	C	$C = \frac{Q}{E}$	−1	−2	2	2
Magnetomotive force	$\mathfrak{F}$	$\mathfrak{F} = \int \mathcal{H} \cdot dl$	0	0	−1	1
Reluctance	$\mathcal{R}$	$\mathcal{R} = \frac{\mathfrak{F}}{\phi}$	−1	−2	0	2
Permeance	$\mathcal{P}$	$\mathcal{P} = \frac{1}{\mathcal{R}}$	1	2	0	−2
Current	I, i	$I = \frac{dQ}{dt} + K' \frac{d\psi}{dt}$	0	0	−1	1
Current density	s	$s = \lim_{\Delta s \to 0} \frac{\Delta I}{\Delta s}$	0	−2	−1	1
Resistance	R, r	$R = \frac{E}{I}$	1	2	−1	−2
Resistivity	ρ	$\rho = \frac{\mathcal{E}}{s}$	1	3	−1	−2
Conductance	G, g	$G = \frac{I}{E} = \frac{1}{R}$	−1	−2	1	2
Conductivity	γ	$\gamma = \frac{s}{\mathcal{E}} = \frac{1}{\rho}$	−1	−3	1	2
Self-inductance	L	$L = \frac{\phi}{I}$	1	2	0	−2
Mutual inductance	M	$M_{AB} = \frac{\phi_A}{I_B}$	1	2	0	−2
Energy	W	$W = EIt$	1	2	−2	0
Power	P	$P = EI$	1	2	−3	0

Table 14. Electrical Units and Dimensions. *(Continued)*

Mks unit	Cgs electrostatic unit*	Cgs electromagnetic unit	Treated in Sec.
1 coulomb	$10^{-1}c(\cong 3 \times 10^{9})$ statcoulombs	10^{-1} abcoulomb	89
$1 \frac{\text{coulomb}}{\text{m}^2}$	$10^{-5}c(\cong 3 \times 10^{5}) \frac{\text{statcoulombs}}{\text{cm}^2}$	$10^{-5} \frac{\text{abcoulomb}}{\text{cm}^2}$	92
$1 \frac{\text{coulomb}}{\text{m}^3}$	$10^{-7}c(\cong 3 \times 10^{3}) \frac{\text{statcoulombs}}{\text{cm}^3}$	$10^{-7} \frac{\text{abcoulomb}}{\text{cm}^3}$	92
$1 \frac{\text{volt}}{\text{m}}$	$10^{6}c^{-1}(\cong \frac{1}{3} \times 10^{-4}) \frac{\text{statvolt}}{\text{cm}}$	$10^{6} \frac{\text{abvolts}}{\text{cm}}$	93
$K'' \frac{\text{coulomb}}{\text{m}^2}$	$K''10^{-5}c(\cong K'' 3 \times 10^{5}) \frac{\text{statvolts}}{\text{cm}}$	$K''10^{-5} \frac{\text{abvolt}}{\text{cm}}$	99
$1 \frac{\text{farad}}{\text{m}}$	$10^{-11}c^{2}(\cong 9 \times 10^{9})$	10^{-11}	99
K'' coulomb	$K''10^{-1}c(\cong K'' 3 \times 10^{9})$ statcoulombs	$K''10^{-1}$ abcoulomb	98
1	$10^{8}c^{-1}(\cong \frac{1}{3} \times 10^{-2})$	10^{8} poles	107, 108
$K'' \frac{\text{amp-turn}}{\text{m}}$	$K''10^{-3}c(\cong K''3 \times 10^{7})$ statoersteds	$K''10^{-3}$ oersted	109, 110
$1 \frac{\text{weber}}{\text{m}^2}$	$10^{4}c(\cong \frac{1}{3} \times 10^{-6}) \frac{\text{statweber}}{\text{cm}^2}$	10^{4} gausses	112
1 weber	$10^{8}c^{-1}(\cong \frac{1}{3} \times 10^{-2})$ statweber	10^{8} maxwells	111
$1 \frac{\text{henry}}{\text{m}}$	$10^{7}c^{-2}(\cong \frac{1}{9} \times 10^{-13})$	$10^{7} \frac{\text{gausses}}{\text{oersted}}$	114, 162
$1 \frac{\text{m}}{\text{henry}}$	$10^{-7}c^{2}(\cong 9 \times 10^{13})$	$10^{-7} \frac{\text{oersted}}{\text{gauss}}$	114, 162
1 volt	$10^{8}c^{-1}(\cong \frac{1}{3} \times 10^{-2})$ statvolt	10^{8} abvolts	96, 129, 130
1 farad	$10^{-9}c^{2}(\cong 9 \times 10^{11})$ statfarads	10^{-9} farad	136
K'' amp-turn	$K''10^{-1}c(\cong K''3 \times 10^{9})$	$K''10^{-1}$ gilbert	126
$K'' \frac{\text{amp-turn}}{\text{weber}}$	$K''10^{-9}c^{2}(\cong K''9 \times 10^{11})$	$K''10^{-9} \frac{\text{gilbert}}{\text{maxwell}}$	114, 162
$K' \frac{\text{weber}}{\text{amp-turn}}$	$K'10^{9}c^{-2}(\cong K'\frac{1}{9} \times 10^{-11})$	$K'10^{9} \frac{\text{maxwells}}{\text{gilbert}}$	162
1 amp	$10^{-1}c(\cong 3 \times 10^{9})$ statamp	10^{-1} abamp	116
$1 \frac{\text{amp}}{\text{m}^2}$	$10^{-5}c(\cong 3 \times 10^{5}) \frac{\text{statamp}}{\text{cm}^2}$	$10^{-5} \frac{\text{abamp}}{\text{cm}^2}$	117
1 ohm	$10^{9}c^{-2}(\cong \frac{1}{9} \times 10^{-11})$ statohm	10^{9} abohms	135
1 ohm-m	$10^{11}c^{-2}(\cong \frac{1}{9} \times 10^{9})$ statohm-cm	10^{11} abohms	119, 135
1 mho	$10^{-9}c^{2}(\cong 9 \times 10^{11})$ statmhos	10^{-9} abmho	135
$1 \frac{\text{mho}}{\text{m}}$	$10^{-11}c^{2}(\cong 9 \times 10^{9}) \frac{\text{statmhos}}{\text{cm}}$	$10^{-11} \frac{\text{abmho}}{\text{cm}}$	119, 135
1 henry	$10^{9}c^{-2}(\cong \frac{1}{9} \times 10^{-11})$ stathenry	10^{9} abhenrys	137, 138
1 henry	$10^{9}c^{-2}(\cong \frac{1}{9} \times 10^{-11})$ stathenry	10^{9} abhenrys	137, 139
1 joule	10^{7} ergs	10^{7} ergs	141, 142, 165
1 watt	$10^{7} \frac{\text{ergs}}{\text{sec}}$	$10^{7} \frac{\text{ergs}}{\text{sec}}$	140

* $c = 2.99796 \times 10^{10}$ cm/sec.

171. Fundamental Units and Dimensions. All engineering quantities may be expressed in terms of a comparatively few quantities that are basic or fundamental. In mechanical units, the fundamental quantities are length, mass, time, and angle. In general, each of these dimensions has a different unit or reference standard in the metric system than in the English system. The name of the unit of measurement is usually sufficient to designate the system as well as the unit of measurement in question. Electrical quantities are usually completely specified in terms of three mechanical units and one electrical quantity (although other combinations are possible) so that mechanical units provide a basis for understanding the fundamental properties of electrical quantities.

For quantities that are derived from, or are more complicated than, the elementary or fundamental quantities, the term "dimension" is used in a slightly different sense than its usual one. In the case of derived units, each new physical quantity gives rise to a new dimension or to a new combination of fundamental quantities to express the new physical quantity. For example, in terms of the fundamental quantities of mass, length, time, and angle, area may be expressed as the product of length and length, so that the dimensions of area are length squared, or, more compactly, $s = l^2$.

Table 13. Mechanical Units and Dimensions

Name	Symbol	Dimensions $m^A l^B t^C \theta^D$				Equivalent magnitudes		
		A	B	C	D	Mks units	Cgs units	English units
Length	l	0	1	0	0	1 m	100 cm	3.281 ft
Mass	m	1	0	0	0	1 kg	1,000 g	2.2046 lb
Time	t	0	0	1	0	1 sec	1 sec	1 sec
Angle	θ	0	0	0	1	1 radian	1 radian	1 radian
Area	s	0	2	0	0	1 sq m	10,000 sq cm	10.76 sq ft
Volume	V	0	3	0	0	1 cu m	10^6 cu cm	35.31 cu ft
Velocity	v	0	1	−1	0	1 m/sec	100 cm/sec	3.281 ft/sec
Acceleration	a	0	1	−2	0	1 m/sec²	100 cm/sec²	3.281 ft/sec²
Density	..	1	−2	0	0	1 kg/m²	0.1 g/cm²	0.2048 lb/ft²
Force	f	1	1	−2	0	1 newton	10^5 dynes	723.30 poundals
Pressure	p	1	−1	−2	0	1 newton/m²	10^3 dynes/cm²	67.222 poundals/ft²
Work (energy)	w	1	2	−2	0	1 joule	10^7 ergs	0.7376 ft-lb
Power	P	1	2	−3	0	1 watt	10^7 ergs/sec	0.7376 ft-lb/sec
Momentum	mv	1	1	−1	0	1 kg-m/sec	10^5 g-cm/sec	7.233 lb-ft/sec
Angular velocity	..	0	0	−1	1	1 radian/sec	1 radian/sec	1 radian/sec
Angular acceleration	..	0	0	−2	1	1 radian/sec²	1 radian/sec²	1 radian/sec²

In another example velocity $v = l/t$, or the ratio of the length to the time has the dimensions of lt^{-1}; acceleration has the dimensions lt^{-2}, force mlt^{-3}, and so on.

In Table 13 are given the more important mechanical units and dimensions in the mks, cgs, and English systems.

172. Electrical Units and Dimensions. Electrical energy and power, being primarily thought of as mechanical in their properties, can be expressed in terms of dimensions which are strictly mechanical and which are given in Table 14. Other electrical quantities cannot be expressed in terms of mechanical units alone but require at least one other (electrical) dimension for their specification. The electrical quantity selected for such dimensional analysis may be arbitrarily chosen. Proposals have been made to select, as the fourth unit, that of charge, current, resistance, or the permeability of free space, and perhaps other quantities as well, as recorded in the excellent treatment by Jauncey and Langsdorf.[1] For developing a philosophy relating all derived quantities in terms of fundamental concepts, the charge is the most logical

[1] Jauncey, G. E. M., and A. S. Langsdorf, "MKS Units and Dimensions," Macmillan, 1940.

fourth unit, although resistance is more suitable for laboratory purposes where working standards are to be maintained. Electrical quantities can be expressed in terms of more than one fundamental electrical concept, with a consequent reduction in the number of mechanical quantities that are then required as fundamental. One system that has been proposed is to omit mass as a fundamental quantity in building up a system of electrical dimensions in which the fundamental quantities would be resistance, current, length, and time. No matter what four quantities are selected as the basis of electrical dimensions, they may be expressed in metric (cgs or mks) or English units; they may also be rationalized or unrationalized depending upon whether the factor 4π (resulting from a fundamental property of space) is missing from expressions for force or potential or whether it is missing from expressions for flux, magnetomotive force, and the field equations.

Table 14 lists the more important electrical quantities together with their dimensions and method of expression in several systems of measurement. The mks system of units is regarded as basic since this system has been standardized or established by international agreement and has the further advantage that its units are, in most cases, the units of practical engineering use. The question of rationalization has not yet been definitely settled, but by using the factors of Table 6 (Sec. 94) the relations of Table 14 apply for either the rationalized or the unrationalized system.

173. Comments on Units and Dimensions. The system of units and dimensions encountered in electrical work is apt to be quite confusing because of the numerous systems of measurement and the various proposals that have been made from time to time. The mks system is the only one that meets the requirements of the scientific worker and the practical engineer alike, and since it has been adopted by international agreement, it appears destined to replace the older cgs systems. The mks system of units eliminates the confusion introduced by the constant of proportionality (numerically equal to the velocity of light in free space), which relates the cgs electrostatic and the cgs electromagnetic systems of units.

The question of electrical dimensions and units is an extensive one and cannot be adequately treated here.

References

"American Standard Definitions of Electrical Terms," AIEE, 1942.

Curtis, H. L.: "Electrical Measurements," McGraw-Hill, 1937.

Gray, A.: "Absolute Measurements in Electricity and Magnetism," Macmillan, 1888.

Hague, B.: "Alternating Current Bridge Methods," Pitman, 1943.

Hund, A.: "High-frequency Measurements," McGraw-Hill, 1933.

Jauncey, G. E. M., and A. S. Langsdorf: "MKS Units and Dimensions," Macmillan, 1940.

Lanchester, F. W.: "Theory of Dimensions and Its Applications for Engineers," Crosby, Lockwood & Sons, London, 1936.

Laws, F. A.: "Electrical Measurements," McGraw-Hill, 1938.

"Radio Instruments and Measurements," *Nat. Bur. Standards Circ.* 74, 1937.

Schulz, E. H., and L. T. Anderson: "Experiments in Electronics and Communication Engineering," Harper, 1943.

Vigoureaux, P., and C. E. Webb: "Principles of Electric and Magnetic Measurements," Blackie, Glasgow, 1936.

CHAPTER 2

RESISTANCE

By Jesse Marsten[1]

1. General Concepts. In any electrical conductor or system in which there is a flow of current, there is a certain amount of energy continually being lost or converted into forms not readily returned to the circuit as electrical energy. As far as is known at present, this dissipation of energy may take one of two forms. There may be an evolution of heat or radiation of energy into space. Such energy dissipation is attributed to a property of electrical conductors or systems termed "resistance."

When dealing with continuous currents, the resistance of a conductor or network R is adequately defined by *Ohm's law*,

$$E = iR \qquad (1)$$

where E is the voltage drop across the conductor or network and i is the current through it. This assumes no back emf due to polarization or other causes. In this case the dissipation of energy takes place entirely in the form of heat generation, and the rate at which electrical energy is thus converted into heat is given by *Joule's law*,

$$P = i^2R \qquad (2)$$

where P is the power or rate at which electrical energy is being dissipated in the form of heat, i is the continuous current in the circuit, and R is the resistance of the circuit.

Ohm's law is insufficient to define resistance in a-c circuits. It is found experimentally that the rate at which energy is lost in a circuit exceeds that which would be necessitated by the resistance of the circuit as determined by Ohm's law. This is because the electromagnetic and electrostatic fields around the circuit vary with time and introduce effects which increase the losses in the circuit. Among these effects may be enumerated the following major ones:

1. *Eddy-current* losses in conductors and other masses of metals in and near the circuit.
2. *Hysteresis* losses in magnetic materials.
3. *Dielectric* losses in the insulating mediums.
4. *Absorption* of energy by neighboring conductors or circuits by induction.
5. *Radiation* of electromagnetic energy into space.
6. *Skin Effect.* Increase of conductor resistance due to nonuniform current density.

All these effects result in an increase in energy loss in the circuit over and above that given by Ohm's law. It therefore becomes necessary to introduce the concept of *a-c resistance* or *effective* resistance, which is defined by the more general Joulean relationship

$$P = i^2R \text{ effective} \qquad (3)$$

where P is the power loss in the circuit due to all causes and i is the effective current in the circuit. Ohm's law for continuous currents follows directly from this more general definition.

2. Units of Resistance. The practical unit of resistance is the *ohm* and is defined by Ohm's law when the voltage and current are unity in the practical system. In 1948, the "international" electrical units, which raised some minor questions as to

[1] International Resistance Co. The author wishes to acknowledge the assistance of Alexander L. Pugh, Jr., and Charles Wellard.

this "unity," were abandoned and a new system of measurement using *absolute* units was officially instituted. Now Ohm's law is correct by definition. Precision resistors are now marked "abs. ohms" to distinguish the standards of measurement. One mean international ohm equals 1.00049 absolute ohms. (Do not confuse with *abohm.*) Owing to the increasing use of resistors having resistances of the order of millions of ohms, the *megohm* unit is also employed. The megohm is equal to 10^6 ohms.

3. Specific Resistance. It is found experimentally that the resistance of an electrical conductor to continuous current is directly proportional to its length and inversely proportional to its cross section:

$$R = \rho \frac{l}{A} \tag{4}$$

The proportionality factor ρ is called the *specific resistance* of the conductor and is a function of the material of the conductor and temperature.

From this definition of specific resistance it is apparent that any number of units may be derived for specific resistance, depending upon the units chosen for l and A. The unit generally employed in practical engineering is *ohms per circular mil foot*, and is the resistance of a 1-ft length of the conductor having a section of 1 cir mil (diameter 1 mil, 0.001 in., for a circular conductor).

Therefore, the resistance of a piece of wire of circular cross section is equal to the ohms per circular mil foot of the metal, times the length in feet divided by the square of the diameter in mils.

4. Volume Resistivity. If, in the above definition, l and A are both unity in the same system of units, then ρ is the resistance between the opposite faces of a unit cube of the material and may be defined as the *volume resistivity* of the material. It should be noted that volume resistivity is not the resistance of any unit volume of the material but is specifically the resistance of unit volume measured between opposite faces whose areas are each unity.

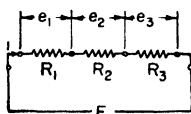



FIG. 1. Simple series circuit.

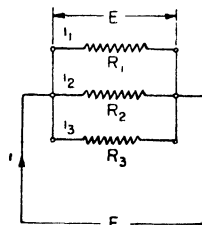



FIG. 2. Parallel circuit.

With a knowledge of the dimensions of a conductor and its specific resistance the resistance of the conductor to d.c. may be computed from Eq. (4). Consistent units must be employed. The resistance thus computed will be correct at the temperature for which the specific resistance applies. To obtain the resistance of the conductor at any other temperature a correction must be applied.

5. Temperature Coefficient. The resistance of a conductor is a function not only of the material and dimensions of the conductor but also of its temperature. Within the temperature limits generally encountered in practice, the change in resistance due to temperature variation is directly proportional to the change in temperature:

$$R_{t2} = R_{t1}[1 + \alpha(t_2 - t_1)] \tag{5}$$

R_{t1} and R_{t2} are the conductor resistances at temperatures t_1 and t_2, respectively.

The proportionality factor α is defined as the *temperature coefficient of resistance* of the material and is the change in resistance of any material per ohm per degree rise in temperature. Sometimes α is multiplied by one million and the temperature coefficient of resistance is reported as ppm/1°C (parts per million per degree centigrade).

All conductors do not react alike to changes in temperature. Metals, for example, have a positive temperature coefficient. Some alloys, such as manganin and constantan, have practically zero temperature coefficient and are therefore used primarily for resistance standards.

A knowledge of the temperature coefficient of conductor materials enables one at times to make more accurate determinations of temperature change than is possible by thermometer measurements, especially in cases where parts to be measured are not readily accessible. Resistance determinations of the conductor are made at two temperatures and the temperature change is computed from Eq. (5).

6. Properties of Materials as Conductors.

Materials	Specific resistance, ohms/cir mil ft	Temperature coefficient α per °C
Pure metals:		
Silver	9.796	+0.0038
Copper	10.37	+0.0039
Gold	14.55	+0.0034
Aluminum	16.06	+0.00446
Tungsten	33.22	+0.0045
Iron (pure)	60.14	+0.0050
Nickel	60	+0.0050
Platinum	63.8	+0.0030
Alloys:		
"30" alloy, alloy 30 (Cu 97.75%, Ni 2.25%)	30.0	+0.0015
"60" alloy, Lohm (Cu 94%, Ni 6%)	60.0	+0.0008
"90" alloy, 95 alloy (Cu 89%, Ni 11%)	90	+0.00049
Hytemco (Ni 72%, Fe 28%)	120	+0.0045
180 alloy, Midohm (Cu 77%, Ni 23%)	180	+0.00018
Manganin (Cu 87%, Mn 13%)	290	±0.000015
Advance, alloy "45," Copel, Cupron, Constantan (Cu 57%, Ni 43%)	294	±0.00002
Nichrome V, alloy "A," Chromel "A," Tophet "A" (Ni 80%, Cr 20%)	650	+0.00011
Nichrome, alloy "C," Chromel "C," Tophet "C" (Ni 61%, Cr 15%, Fe 24%)	675	+0.00015
Radiohm (Fe 78.5%, Cr 16.5%, Al 5%)	800	+0.00007
"331" alloy, Karma, Evanohm (Ni, Cr, Al, and small trace of Fe)	800	±0.00002

7. Resistors in Series and Parallel. Simple and complex networks of resistors may be represented by an equivalent resistor expressed in terms of the individual resistances making up the network. The equivalent resistance of a number of resistors connected in series is equal to the sum of the individual resistances. Referring to Fig. 1,

$$E = iR_{\text{equiv}} = e_1 + e_2 + \cdots + e_n = R_1 i + R_2 i + \cdots + R_n i = i(R_1 + R_2 + \cdots + R_n)$$

$$\frac{E}{i} = R_{\text{equiv}} = (R_1 + R_2 + \cdots + R_n)$$

$$R_{\text{equiv}} = \sum_1^n R$$

The reciprocal of the equivalent resistance of a number of resistors connected in parallel is equal to the sum of the reciprocals of the individual resistances. Referring to Fig. 2,

$$i = i_1 + i_2 + \cdots + i_n = \frac{E}{R_1} + \frac{E}{R_2} + \cdots + \frac{E}{R_n}$$

$$\frac{i}{E} = \frac{1}{R_{\text{equiv}}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}$$

$$\frac{1}{R_{\text{equiv}}} = \sum_1^n \frac{1}{R}$$

RESISTANCE AS FUNCTION OF FREQUENCY

8. Ratio of H-f Resistance to the D-c Resistance for Different Values of $x = \pi d \sqrt{2\mu f/\rho} \times \sqrt{1/1{,}000}$.

x	R_f/R_0	x	R_f/R_0	x	R_f/R_0
0	1.0000	5.2	2.114	14.0	5.209
0.5	1.0003	5.4	2.184	14.5	5.386
0.6	1.0007	5.6	2.254	15.0	5.562
0.7	1.0012	5.8	2.324		
0.8	1.0021	6.0	2.394	16.0	5.915
0.9	1.0034	6.2	2.463	17.0	6.268
				18.0	6.621
1.0	1.005	6.4	2.533	19.0	6.974
1.1	1.008	6.6	2.603	20.0	7.328
1.2	1.011	6.8	2.673		
1.3	1.015	7.0	2.743	21.0	7.681
1.4	1.020	7.2	2.813	22.0	8.034
1.5	1.026	7.4	2.884	23.0	8.387
				24.0	8.741
1.6	1.033	7.6	2.954	25.0	9.094
1.7	1.042	7.8	3.024		
1.8	1.052	8.0	3.094	26.0	9.447
1.9	1.064	8.2	3.165	28.0	10.15
2.0	1.078	8.4	3.235	30.0	10.86
				32.0	11.57
2.2	1.111	8.6	3.306	34.0	12.27
2.4	1.152	8.8	3.376		
2.6	1.201	9.0	3.446	36.0	12.98
2.8	1.256	9.2	3.517	38.0	13.69
3.0	1.318	9.4	3.587	40.0	14.40
				42.0	15.10
3.2	1.385	9.6	3.658	44.0	15.81
3.4	1.456	9.8	3.728		
3.6	1.529	10.0	3.799	46.0	16.52
3.8	1.603	10.5	3.975	48.0	17.22
4.0	1.678	11.0	4.151	50.0	17.93
				60.0	21.47
4.2	1.752	11.5	4.327	70.0	25.00
4.4	1.826	12.0	4.504		
4.6	1.899	12.5	4.680	80.0	28.54
4.8	1.971	13.0	4.856	90.0	32.07
5.0	2.043	13.5	5.033	100.0	35.61

It is frequently useful to know the largest diameter of wire of different materials that will give a ratio of R_f/R_0 of 1.01 for different frequencies. For a ratio of R_f/R_0 equal to 1.001, the diameters given in Sec. 10 should be multiplied by 0.55; and for R_f/R_0 equal to 1.1, the diameters should be multiplied by 1.78.

9. Skin Effect. The resistance of a conductor is a minimum when the current density is uniformly distributed over the cross section of the conductor. This condition occurs with d.c. The resistance increases for nonuniform distribution of current density over the cross section of the conductor. The latter condition exists in conductors carrying a.c. This is a result of the distribution of magnetic flux lines, outside and inside the conductor. If the conductor is assumed to be made up of a number of conducting elements in parallel, then the interior elements, being surrounded by more flux lines than the exterior, will have greater reactance and, therefore, the current in the interior elements will be less than that in the exterior elements. As a result, the current crowds toward the surface of the conductor, giving a nonuniform current density. This imperfect penetration of current in a conductor, resulting in an increase in resistance, is termed "skin effect."

This effect is important in high-Q circuits but is seldom significant in the usual application of high-resistance alloys as small-diameter wires.

Skin effect in a conductor is a function of the following factor:

$$t\sqrt{\frac{\mu f}{\rho}} \qquad (6)$$

where t = thickness of the conductor
f = frequency of current
μ = permeability of the conductor
ρ = specific resistance of the conductor, microhm-cm*

It is possible to compute accurately the h-f resistance of simple round cylindrical conductors from involved functions of the above factor. To facilitate these computations, tables have been prepared from which the ratio of h-f resistance R_f to d-c resistance R_0 may be quickly determined. From this factor and the easily measured d-c resistance, the h-f resistance may be computed.

The table in Sec. 8 gives the values of R_f/R_0 for different values of the factor

$$x = \pi d\sqrt{\frac{2\mu f}{\rho}}\sqrt{\frac{1}{1{,}000}}$$

where d is the diameter of the wire in centimeters, ρ is the volume resistivity in microhm-centimeters (1.724 at 10°C for copper), x may be computed for any particular case, and R_0 may be measured at d.c. or computed.

11. Reduction of Skin Effect. In view of the tendency of the current to crowd to the surface of the conductor at high frequencies, the remedies that have been found practical in effecting an improvement in the resistance ratio R_f/R_0 have been those in which the conductor has been designed so that it presents a thin skin to the current flow; *i.e.*, the conductor should be a thin skin rather than a solid conductor of large cross section of which only a small part is used at high frequencies. These are:

1. *Use of Flat Copper Strip.* While skin effect is present, for the same cross-sectional area a flat strip gives a lower resistance ratio than do round conductors.

* Microhms per centimeter length of a conductor having a cross section of 1 sq cm.

10. Maximum Diameter of Wires for H-f Resistance Ratio of 1.01.

Frequency, kc	100	400	1,000	1,600	2,000	3,000	6 Mc	10 Mc	20 Mc	60 Mc	300 Mc
Wave length, m	3,000	750	300	187.5	150	100	50	30	15	5	1
Material	Diameter, cm										
Copper	0.0356	0.0177	0.0112	0.0089	0.0079	0.0065	0.00457	0.00355	0.00251	0.00145	0.00065
Silver	0.0345	0.0172	0.0109	0.0086	0.0077	0.0063	0.00445	0.00346	0.00244	0.00141	0.00063
Gold	0.0420	0.0210	0.0133	0.0105	0.0094	0.0077	0.00543	0.00422	0.00298	0.00172	0.00077
Platinum	0.1120	0.0560	0.0354	0.0280	0.0250	0.0205	0.01445	0.0112	0.00783	0.00456	0.00205
Mercury	0.264	0.132	0.0836	0.0661	0.0591	0.0483	0.03416	0.0265	0.0187	0.0108	0.00485
Manganin	0.1784	0.0892	0.0564	0.0446	0.0399	0.0325	0.02300	0.0179	0.0126	0.0073	0.00326
Constantan	0.1892	0.0946	0.0598	0.0473	0.0423	0.0345	0.02440	0.019	0.0134	0.00775	0.00346
German silver	0.1942	0.0970	0.0614	0.0485	0.0434	0.0354	0.02500	0.0195	0.0138	0.00784	0.00354
Graphite	0.765	0.383	0.242	0.191	0.171	0.140	0.0988	0.0767	0.0542	0.0312	0.0140
Carbon	1.60	0.801	0.506	0.400	0.358	0.292	0.2065	0.16	0.1135	0.0655	0.0292
Iron:											
μ = 1,000	0.00263	0.00131	0.00083	0.00066	0.00059	0.00048	0.000339	0.000263	0.000186	0.000107	0.000048
μ = 500	0.00373	0.00187	0.00118	0.00093	0.00084	0.00068	0.00049	0.000374	0.000264	0.000152	0.000068
μ = 100	0.00838	0.00418	0.00264	0.00209	0.00186	0.00152	0.001075	0.000836	0.000591	0.000341	0.000153

2. *Use of Tubular Conductors.* The resistance of the internal portion of the conductor is affected much more by increase in frequency than the outer portion; therefore, the resistance of thin tubular conductors changes less with increase in frequency than does that of solid conductors.

3. *Use of Litzendraht.* According to Eq. (6), the smaller the diameter of the wire, the less the skin effect. Litzendraht is a braided cable made up of a large number of fine strands of wire. When certain precautions are taken, this braid shows a very much lower resistance ratio than does a solid copper wire of equal section. These precautions are:

a. Each strand must be thoroughly insulated from every other strand to avoid contact resistance.

b. Braiding must be such that each strand passes from the center to the outside of the conductor at regular intervals—a sort of transposition. This ensures that all strands are affected alike by the magnetic flux.

c. Each strand must be continuous.

GENERAL CONSTRUCTION OF RESISTORS

Basic elements in the construction of resistors are:

1. Base
2. Resistive element
3. Terminations
4. Protective cover

12. Base. Important characteristics of a base suitable for the manufacture of resistors are:

1. It should have high insulation resistance at room and elevated temperatures.
2. It should not have much ionizable content; otherwise ionic conduction will result with consequent electrolysis and corrosive effect.
3. It should have high mechanical strength.
4. It should have low moisture absorption; otherwise it should be protected against moisture.
5. It should be capable of withstanding high thermal shock.
6. Its coefficient of thermal expansion should approximate as closely as possible that of the resistance element and protective coating of the resistor.
7. It should have low h-f losses when the resistor is intended for high frequencies.

The more common choices for resistor bases are:

1. Ceramic
2. Glass
3. Plastics

Ceramics. Generally used for winding forms for precision and power wire-wound resistors. They are used for the base of high-frequency and high-voltage composition resistors, for pyrolitically deposited carbon, boron-carbon, and for metal film resistors. The ceramic properties, which vary with composition and with details of their manufacture, influence the electrical properties of the resistor.

Glass. Used as the insulating base in one of the types of fixed composition resistors. It is used as the base for conductive-oxide types of resistors, and for the base for some metallic-film types. In the form of fiber glass, it is formed into a hard, somewhat flexible bundle and used as the base for some wire-wound resistors. Glass offers the same general advantages as ceramic; it has a smoother surface but is more subject to damage by shock unless reinforced by means such as molding.

Plastics. Used as the base in composition variable resistors. Plastics or fiber is used for the base of many wire-wound resistors of the variable and the molded types and also precision wire-wound resistors. Plastics may not be operated at as high temperatures as ceramics or glass but they are not so brittle and may be die-cut and formed readily.

13. Resistive Element. The more common types of material used for the resistive element are:

1. Metal alloys (wire)
2. Carbon and graphite, usually with binders
3. Pyrolitically deposited films
4. Metallic films
5. Conductive oxides

14. Metal Alloys. The resistive properties of the more common metals and alloys used for resistors are given in Sec. 6. These properties are for bulk materials and do not hold if a dimension of the resistor element is the order of the mean free path of the electrons (see Sec. 17). Generally, alloys have higher specific resistance and lower temperature coefficients of resistance than pure metals. For these reasons, alloys are generally used for wire-wound resistors.

Three classes of wire, which may be round or flat, are generally used:

Nickel-Copper. Some of the more commonly known trade names are Advance, Constantan, Cupron, Midohm, and 30 Alloy. The nickel-copper alloys are made to cover a wide range of specific resistivities, which makes it possible to make very low resistance values without resorting to excessively large wire and resistor sizes. The temperature coefficient of these alloys, however, increases as the specific resistivity decreases. Constantan (Advance, Cupron) has a negligible temperature coefficient and is therefore widely used for precision resistors as well as in other types of resistors. It has a high thermal emf against copper and iron and is therefore widely used in thermocouples.

Nickel-Chromium. Commonly known by the trade names of Nichrome, Tophet, and Chromel. This class of alloys has a very low temperature coefficient and high specific resistance. It is made in two formulations: one iron-free, the other with a substantial percentage of iron.

Nickel-Chromium-Aluminum. Commonly known by the trade names of 331 Alloy, Evanohm, and Karma. This class of alloys has a very low temperature coefficient and the highest specific resistance of those commercially useful in resistors.

Alloy wires are made in sizes as low as 0.0005 in. in diameter. However, in resistor design there is usually a limit placed on the minimum allowable size. This is usually set at 0.001 in., the object being to reduce the possibility of failure due to corrosion and mechanical weakness. Where nickel-chromium alloy is used, it is sometimes recommended that the iron-free wire be used for wire sizes smaller than 0.002 in.

15. Carbon and Graphite. These conductors are used usually with binders to form the resistance element in the so-called composition-type resistors, both fixed and variable. The term *composition-type resistor* is employed to cover that group of resistors in which a conductor is mixed with a binder in definite proportions and suitably treated to produce a resistor material. The physical form it takes may be a solid body or a film. By and large, these resistors, whether of the solid-body or film type, are made in the same general way, although there are wide variations in materials, formulations, details of manufacturing procedure, etc. Resistor elements of this class are made by preparing a homogeneous mix of a conductor, generally graphite or carbon or both, with a resinous binder and sometimes a filler. In the case of the solid resistor, the binder is a resinous molding powder; in the case of the film resistor, the resin is in the form of a varnish. In both cases a curing operation is essential to harden the mass or film. The mixtures have infinite or very high resistance prior to the final polymerizing operation. This operation involving temperature, pressure, or both shrinks and fuses the resistive composition into a tough hard body or film. The resulting structure then consists of conducting particles held together by a dielectric binder. Resistance values are controlled by modifying the ratio of dielectric to conductor. The higher this ratio, the higher the resistance value. The conducting particles make contact with each other under the pressure that exists in the structure. Most of the resistivity of the device is a summation of the contact resistances between particles, so that whatever alters the contact pressure will affect the resistivity. Many of the performance characteristics of this class of resistor are a direct result of the structure of this device.

16. Pyrolitically Deposited Films. These resistive elements are made by decomposing hydrocarbons or other materials at very high temperatures to form conductive films upon a heat-resistant base, usually ceramic. The deposits are usually carbon but may be mixtures of carbon and other materials such as boron. The electrical properties are very good over a relatively narrow band of specific resistance; therefore the higher values are usually made by helixing (frequently called *spiraling*) the resistance film to the desired resistance values.

17. Metallic Films. These resistive elements are made by the decomposition of metallic compounds at high temperature forming metallic films, or by the evaporation or sputtering of metals upon a base under vacuum. They may be made from noble metals, alloys, or pure metals. These films are usually made very thin and it has been found that thin films have electrical properties that differ from those of the bulk material from which they are made. The specific resistance is much higher and the temperature coefficient of resistance varies from plus to minus, depending upon the material, the film thickness, and other factors. By proper selection of the materials, thickness of film, methods of deposition, and helixing, resistors of a wide range of resistance values are made that have temperature coefficients (T.C.) within 25 ppm per °C of any desired value between plus or minus 200 ppm (the mean value of T.C. can be more positive or negative on special orders). The films are very stable with time, with both low and high temperature, and withstand moisture and abrasion.

18. Conductive Oxide. Conductive-oxide resistive elements are made by fusing certain metallic oxides into the surface of glasses. The usual materials are combinations of tin and antimony oxides on pyrex glass which yield resistors of intermediate temperature coefficient of resistance and of intermediate resistance values.

19. Terminations. Terminations of resistors fall into four classes:

1. Wire leads
2. Lugs
3. Ferrules
4. Special

1. *Wire Leads, Radial or Axial.* Used on composition resistors, many lightweight resistors of the precision, low-power wire-wound, pyrolitic film, metallic-film, conductive-oxide types, and on some printed circuits. They are usually of solid tinned copper wire of a size suitable for the weight of the resistor. Some leads are coated with alloys that permit easy soldering even after long storage before assembly.

2. *Lug Terminals.* Used on variable resistors, power wire-wound resistors, precision resistors, and some others. They are usually of tinned alloys. Although most are supplied with holes to take wire some are "blind" for spot welding the resistors into the circuit.

3. *Ferrules.* Usually of a size to fit a standard cylindrical fuse clip and are made of alloys. They are usually used on high-power wire-wound resistors and for meter-multiplier resistors.

4. A few resistors are made with special terminals unique for their application such as for use in coaxial lines, and some, which are used as attenuators in wave guides, have no terminals.

20. Protective Cover. Most resistor elements are insulated and protected from damage during shipment, assembly, in use, and against adverse environment by a protective coating. The coatings may be nothing in the case of high-power water-cooled resistors where excellent heat transfer is essential but may be a thick plastic in some cases.

The usual protective coatings are:

1. Vitreous-enamel coatings
2. Cement coatings employing inorganic binders
3. Ceramic tube enclosures
4. Glass or ceramic tube enclosures, hermetically sealed
5. Cement coatings employing organic binders
6. Molded or cast plastic enclosures
7. Films

Coverings in the first four classifications are capable of withstanding temperatures in excess of 250°C without deterioration. They afford a high measure of protection against humidity. Exceptions to this statement are coverings employing sodium silicate (water-glass) binders which are highly hygroscopic and therefore unsuitable where resistance to humidity is an important factor.

The hermetically sealed tube enclosure affords the maximum protection against humidity. Such resistors are capable of standing severe thermal shock and numerous alternate immersions in salt water solution at 0 and 100°C, a requirement often specified by the U.S. Navy.

Coverings in the fifth classification are capable of withstanding temperatures up to about 175°C, varying with the nature of the binder. Resinous binders in general stand lower temperatures than asphaltic binders. They are, however, superior to the high-temperature coatings in their moisture-resistant properties. Silicone resin binders are capable of withstanding temperatures as high as 200°C and afford a high measure of protection against humidity.

Coverings of molded or cast plastic are capable of withstanding temperatures from 100 to 160°C, depending upon the nature of the molding material used. The ordinary general-purpose molding materials with wood-flour base are good at the lower temperatures, whereas the asbestos and mica-filled materials are good at the higher temperatures.

Coverings of the last classification are of one or more films of appropriate varnishes sometimes filled with a pigment to improve abrasion resistance. When carefully selected, these films provide some protection against humidity and mechanical damage.

RESISTOR RATINGS

21. Power Rating of Resistors. The power rating is primarily determined by the power that the resistor can dissipate continuously for an unlimited time without exceeding a permissible maximum temperature and without exceeding arbitrarily chosen performance limits. This limits the permissible temperature rise, which varies for different classes of resistors. The temperature rise is measured from a specified ambient temperature. Ambient temperature is defined as the temperature of the surrounding air.

22. Temperature Rise of Resistors. The temperature rise of resistors varies with the design, cooling, and power load. The temperature rise varies somewhat with different makes of resistors owing to differences in coating materials such as heat conductivity, texture, and color.

The 100 per cent rating is based upon the specified maximum temperature rise in free air. Resistors made to the same specifications have similar temperature rises, independent of resistance, when the rises for the same powers are compared. For this reason, typical temperature-rise curves, as given in later sections, can be used with reasonable accuracy in the design of equipment.

23. Power Derating. Nominal power ratings are based on specified good ventilation, which is seldom encountered in practice. The actual load at which resistors may be used safely is therefore usually less than the nominal rating. For example, in the case of high-power resistors nominally rated on the basis of 250°C rise from a 25°C ambient temperature, it is generally recommended that when used in confined space the loads be reduced to one-fourth to one-half nominal rating, resulting in temperature rises of 80 to 160°C. In practice, even these temperature rises may be excessive because of the proximity of the resistors to other parts that may not be subjected to elevated temperature, or to fire hazard, underwriters' requirements, etc. The specific application therefore limits the practical use of a resistor rather than the nominal rating.

In practice, the above considerations are taken into account by "derating" resistors. If a resistor is used at its nominal rating at an ambient temperature greater than its reference ambient, the maximum permissible temperature will be exceeded. To avoid

this, the temperature rise must be limited by the use of power low enough to compensate for the increased ambient. This is done by means of derating curves for each class of resistor showing ambient temperature vs. permissible operating load. This principle of derating applies to all classes of resistors. Every class of resistor can have such a derating curve computed from its rating and temperature-rise curve. Figure 3 shows a typical derating curve for fixed composition resistors. For derating curves applicable to specific types of resistors, see the appropriate sections. Derating curves of other resistors may be secured from the appropriate EIA or government specifications (see Sec. 24) or from the various manufacturers.

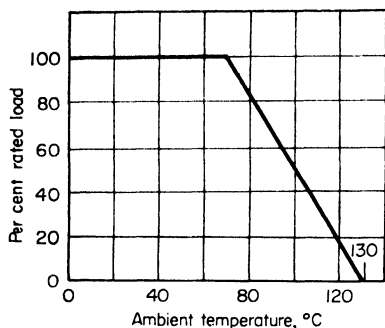



FIG. 3. Typical power derating curve for fixed composition resistors.

FIXED LOW-POWER RESISTORS

The types available include composition and wire-wound types. Each type has unique properties and limitations. These resistors have power ratings up to 2 watts. They are small, up to 1¾ in. long and up to 11/32 in. in diameter. Most are made with axial leads, 1½ in. long, of tinned copper wire of 18 gage or less diameter. The power ratings are usually 1/10, ¼, ½, 1, and 2 watts with maximum voltage ratings of 500 volts or less, depending upon power rating. More resistors of this type are used than all other types combined.

24. Wire-wound Low-power Resistors. The wire-wound types are available from 0.24 to 8,200 ohms but depend upon size. These units dissipate ½ to 1 watt per sq in. of surface. There are two general styles: (1) flexible resistors and (2) plastic-insulated resistors. Both have a core of either cord or glass fiber. The former has a textile or glass-fiber covering; the latter is molded in plastic. They are equipped with wire leads making them very convenient for "point-to-point" wiring in circuits, eliminating the necessity for special fittings for mounting. The flexible resistors with glass-fiber cores and coverings are capable of much higher ratings. The plastic-molded style has become standardized in the radio industry.

The wire-wound types are available in lower resistance values and have lower changes with applied voltage and lower noise levels than the composition types. The inductance and distributed capacitance are slightly greater and more variable, resistor to resistor, than the composition type. They are used extensively where

Table 1. Wire-wound Resistors with Plastic Insulation

Rating, watts based on 40°C ambient	Size, in.		Min resistance, ohms	Approx inductance, μh	Max resistance (JAN-R-184), ohms	Approx inductance, μh	Max resistance (standard), ohms	Approx inductance, μh
	Diam	Length						
½	3/16	5/8	0.24	0.087	470	0.6	870	2.5
1	¼	1¼	0.47	0.17	2,200	8.0	5,100	33
2	5/16	1¾	1.0	0.35	3,300	13	8,200	56

inexpensive, small, low-power resistors with good stability are required as, for example, in audio circuits, particularly with cathode follower, vibrator circuits, telephone circuits, radio, TV, etc.

Table 1 gives typical information about this class of resistor. The military specification covering this class is JAN-R-184; the industry specification is EIA REC-117.

Power Derating. The insulated wire-wound resistors of this class are derated for high ambient temperatures as shown in Fig. 4.

Color Code. The color code for resistance value and resistance tolerance is the same as described in Sec. 34 for composition resistors except that the first, or A, band is twice as wide as the other bands. The wide A band identifies the resistor as of wire-wound construction.

Standard Resistance Values. The standard resistance values per decade are the same as described in Sec. 35 for composition resistors.

FIG. 4. Power derating curve for molded wire-wound resistors (JAN-R-184 and EIA REC-117).

25. Composition Resistors. The composition-type resistor has attained a wide popularity in the manufacture of radio and allied products because of the following advantages: (1) flexibility in resistance—it may be made in any values from 10 ohms up to 100 megohms; (2) compactness—its physical dimensions are small for any resistance, and they may be made in sizes as small as 0.067 in. in diameter by 0.140 in. long.

Numerous styles of these resistors have been produced, but they take two general forms:

1. *Solid-body Resistors.* In this style the resistor material is extruded, pressed, or molded into its final physical form, which generally is a solid rod. After this it may be subjected to some form of heat-treatment. The so-called *carbon* resistor is an example of this style. It is produced both as an uninsulated resistor (as described) and as an insulated resistor with an outside molded insulating covering. Wire leads are used as terminals.
2. *Filament-coated Resistors.* Here a conducting coat or film is baked on the surface of a continuous glass filament or other form. In the case of the glass filament this is completely enclosed in an insulating housing. The so-called *insulated-filament* resistors are examples of this style. Wire leads are used as terminals.

The insulated type of resistor, in which the resistor element is molded in a plastic covering, is used almost universally and may be considered standard in the industry.

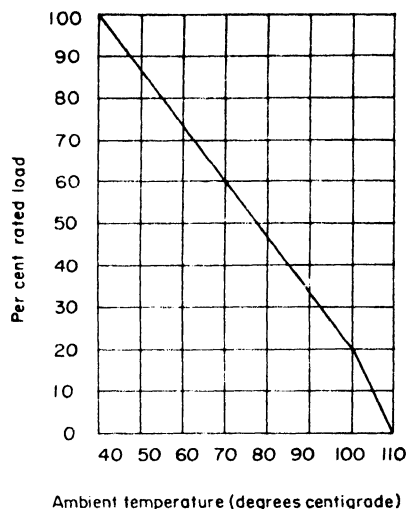
Table 2 gives typical information about this class of resistor. The military specification covering this class is MIL-R-11B. The industry specification, EIA RS 172, includes more sizes and classes.

Table 2. Insulated Composition Resistors

Power rating, watts (70°C ambient)	Max diam of resistor, in.	Max over-all length of resistors, in.	Max rated voltage, volts, d-c or rms
1/10	0.075	0.160	150
1/4	3/32	3/8	250
1/2	3/16	3/8	350
1	1/4	3/4	500
2	5/16	1 5/16	500

Composition resistors possess properties differing from those of metallic resistors. The most important, possessed by all styles in varying degree, are as follows:

26. Voltage Characteristics. The resistance is not independent of the applied voltage and decreases with increasing voltage. This characteristic is measured by a fac-

tor called *voltage coefficient* defined as

$$\text{Voltage coefficient} = 100 \frac{R_1 - R_2}{R_2} \times \frac{1}{E_1 - E_2}$$

where E_1 = rated continuous working voltage (usually applied for 0.1 sec at 10-sec intervals to avoid measuring the change due to temperature rise)
E_2 = 0.1 rated continuous working voltage
R_1 = resistance at rated continuous working voltage
R_2 = resistance at 0.1 rated continuous working voltage

The voltage coefficient is a function of

1. Composition of resistance mix.
2. Resistance value. For a given type of composition or mix and size of resistor, the coefficient decreases with decreasing resistance value.
3. Length of resistor. For a given composition and resistance value, the coefficient decreases with increasing length.

These facts are generally true of all types of composition resistors and are illustrated in Fig. 5, taken on resistors of two different lengths made with two resistance-material formulas. The reason for this behavior will be clear from the following: Lower resistance values for a given composition imply less insulating material in the

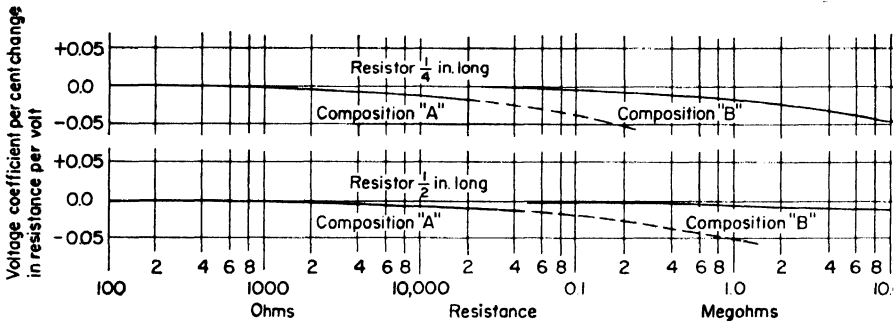



FIG. 5. Voltage characteristics of composition resistors.

mix. Greater length of resistor for a given value implies lower resistance per unit length of resistor and, therefore, less insulating material. The variation required in each of these factors to produce minimum voltage coefficient is in the direction to reduce the amount of the dielectric required in the mix.

The explanation is to be found largely in the observed relationship between the contact resistance of carbon particles and the voltage across contacts. The contact resistance between two carbon particles is an inverse function of voltage. Since most of the resistance resides in the contacts, anything that reduces the voltage across the contact of conducting particles will reduce the voltage coefficient. Increasing the number of conducting particles is the equivalent of reducing the amount of dielectric in the resistor, which is the condition for minimum voltage coefficient.

27. High-frequency Characteristics—Boella Effect. Unlike wire-wound resistors, which generally show a slight increase in resistance with frequency (skin effect), the composition-type resistor shows a pronounced fall-off in apparent a-c resistance from its d-c value with increasing frequency. This is particularly true of high resistance values. This effect, known as the "Boella effect" after its discoverer in Italy, was further investigated in England, where G. W. O. Howe proposed a theory explaining this behavior. He assumed that the resistor simulated a transmission line of half the length of the resistor, with uniformly distributed resistance and capacitance (see Fig. 6). He derived a universal relationship between R_f/R_0 and fR_0, in which the

ratio of R_f/R_0 is constant for equal products of R_0 and f. (R_f is the apparent h-f resistance at frequency f; R_0 equals the d-c resistance.) The fall-off in resistance value with increasing frequency is, therefore, due primarily to the shunting effect of the distributed capacitance in the resistor. Many observations confirm this theory generally, although Salzburg and Miller[1] show considerable departure at frequencies above 200 Mc. This subject was studied by R. F. Field. In an unpublished paper, he introduced the additional consideration of dielectric loss in the distributed and lumped capacitances in Howe's artificial line, which would increase still further the fall-off in apparent a-c resistance. The Boella effect in composition resistors of all types is a direct result of the resistor structure. Large numbers of conducting particles interspersed with a large amount of dielectric produce capacitances and their concomitant dielectric losses, which act as shunts at high frequencies.

FIG. 6. (*a*) Composition resistor of length L represented with lumped parameters. (*b*) Representation of (*a*) as transmission line having a length $L/2$ with uniformly distributed R and C.

28. Conditions for Good H-f Characteristics. On the basis of this analysis two important conditions follow for optimum h-f characteristics:

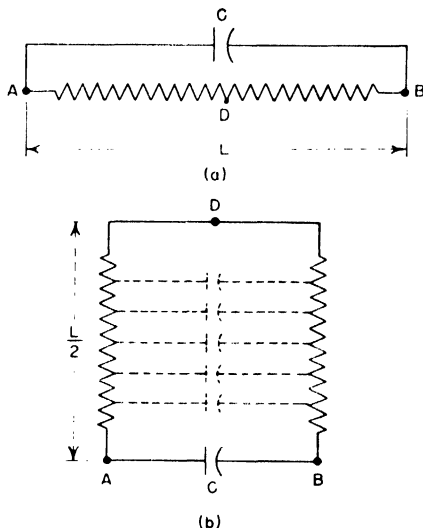
1. *Minimum dielectric* in the resistor mass and surrounding it. The dielectric constant and loss factors should be minimum.
2. *Geometry of Resistor.* As a direct result of this condition, the geometry of the resistor influences the h-f characteristics. Consider two resistors of equal length, one having twice the cross section of the other. For a given resistance value, the smaller unit would require resistance material with the lower specific resistivity. This means less dielectric, resulting in better frequency characteristics. The condition for minimum dielectric is therefore minimum cross section of resistor. Similarly, if the areas of both resistors are constant, the condition for minimum dielectric is maximum length. For best h-f characteristics, the ratio of area to length of the resistor should be a minimum.

The filament-type resistor most nearly meets both requirements established for minimum Boella effect. Its cross section is extremely small since it is a film. (Film thicknesses in commercial practice vary from 0.0002 to 0.001 in.) The film volume also being small, the dielectric mass is very low. Extensive tests of frequency characteristics of composition resistors have recently been made.[2] Curves from these data are shown in Figs. 7 and 8.

The reduction of resistance with increase of frequency of the electric signal is greater with the solid-body type. Most of the other properties are closely the same for the two types.

29. Resistor Noise Characteristics. Composition resistors all show, in varying degree, the presence of microphonic noise. The degree of noise is a function of the voltage, size of the resistor, resistance value, and the nature of the materials used. In general, for a given set of materials in the resistor, the noise level increases with increasing voltage, increasing resistance, and decreasing size.

It is interesting to note that the pattern of factors influencing noise is similar to that influencing the voltage coefficient and h-f characteristics of composition resistors.

[1] *RCA Rev.*, April, 1939.
[2] Interim Report, Part II, "Service Facilities and Materials for Performance Evaluation of Existing Types of Fixed Resistors in Ultra High Frequency Spectrum," Report 19A, Signal Corps Contract No. W 36-039-SC-44526.

Christensen and Pearson[1] reported that they failed to detect any noise in a solid carbon filament other than that due to thermal agitation. As a result, they concluded that most of the noise in carbon-composition resistors came from the contacts between carbon particles and that this was a function of voltage. They demonstrated that increasing the number of contacts would decrease the over-all noise. Empirical results with solid and film types of composition resistors confirm their conclusions.

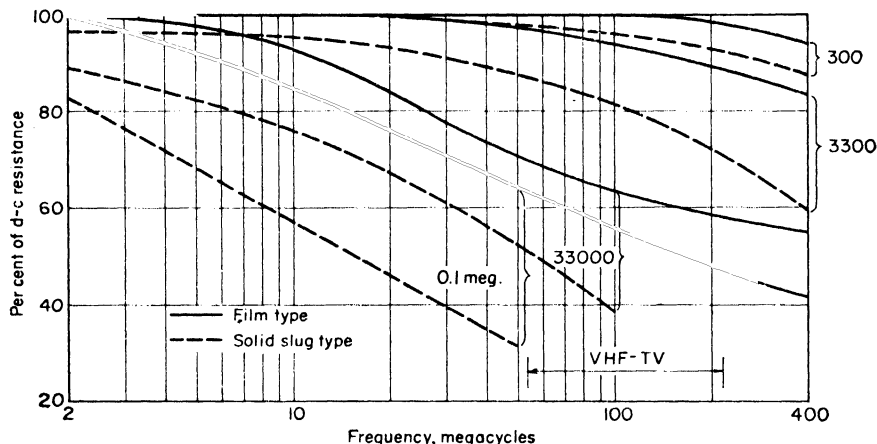



FIG. 7. Frequency characteristic of thin-film vs. solid-slug types of ½-watt resistors.

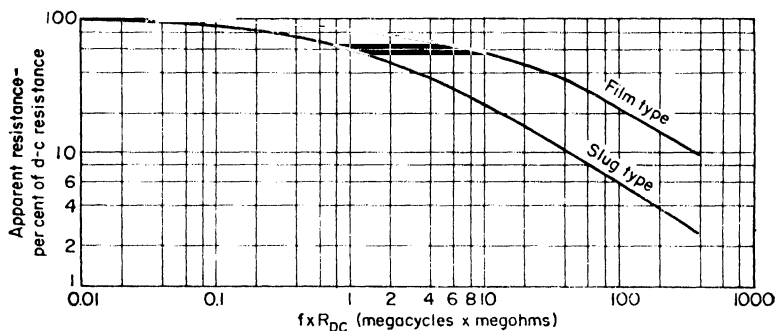



FIG. 8. Composition-resistor frequency characteristics.

In the type of resistor under discussion, resistance is increased by adding dielectric binder. This is equivalent to reducing the contact density or the number of contacts per unit volume and increasing the resistance of the contacts. For a given current through the resistor or a given power in the resistor, this results in greater voltage drop across each contact, which produces greater noise. As in the h-f and voltage characteristic the condition for minimum noise is minimum dielectric in the resistor.

Noise is difficult to measure quantitatively and is not now a specified characteristic of resistors. Most composition-resistor manufacturers are reducing the noise level of their resistors, and most are so low that the microphonic noise cannot be detected in any but highly specialized amplifiers.

30. Humidity Characteristics. Although the general effect of humidity on resistors is to reduce resistance value because of resultant leakage, its effect on composition resistors is the opposite: resistance increases. This is because moisture causes expan-

[1] *Bell System Tech. J.*, April, 1936.

sion of the resistor binder material and separation of the conducting particles, with resulting increase in resistance value. This effect may be reduced by proper choice of materials and by suitable treatment of the resistor.

31. Temperature Characteristics of Composition Resistors. The T.C. of composition resistors is not a constant factor. It varies with resistance value, and for a given resistance it varies with temperature. It may even change sign at some point on the temperature scale.

This anomalous behavior is due to the peculiar structure of the resistor. A composition resistor is one in which the conductor (graphite, carbon, or other material), which may have negative and/or positive coefficients, is mixed in varying proportions with insulation. In addition, there is an insulating supporting base for the film-type

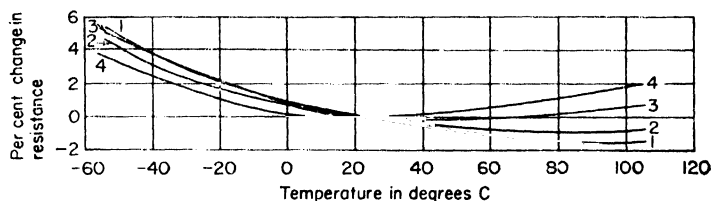



FIG. 9. Resistance-temperature characteristic, ½ watt, ⅛ by ⅜ in., composition resistor, Curve 1, 10 megohms; curve 2, 0.27 megohm; curve 3, 10,000 ohms; curve 4, 1,000 ohms.

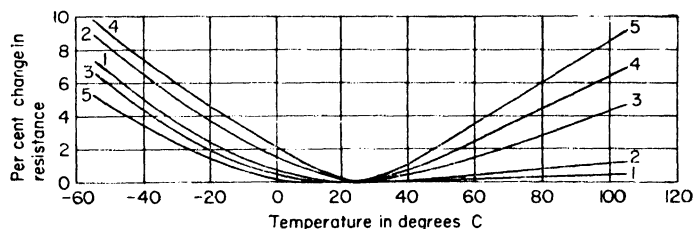



FIG. 10. Resistance-temperature characteristic, ½ watt, ⅛ by ⅜ in., solid composition resistors. Curve 1, 1,000 ohms; curve 2, 10,000 ohms; curve 3, 0.100 megohm; curve 4, 1.00 megohm; curve 5, 10.0 megohms.

resistor. All these components react differently to temperature changes. Stresses are set up by even small differences in coefficient of expansion between these various components. Also, contraction and expansion may affect the contact pressure between conducting particles nonlinearly for the same temperature increment at different temperatures. The T.C. of the conducting material itself—graphite, carbon, or other material—is not the dominant factor in establishing the temperature characteristic of the resistor. Rather it is the result of all the stresses set up by differential expansion, which affects the contact pressure between conducting particles, and this pressure determines the ultimate resistance and therefore the T.C. As a result, it is necessary to speak of temperature characteristics over a wide range of temperatures rather than to cite a constant T.C.

Figures 9 and 10 show typical temperature characteristics of two types of composition resistors for different values and over a wide range of temperatures.

32. Power Rating Composition-type Resistors. The rating of composition-type resistors is complicated by many variables. The T.C. being large, it is possible for a resistance change to become quite appreciable before a temperature limitation is exceeded. Furthermore, with the higher resistances, such as 0.27 megohm and higher, in which the power dissipation may be very low, the voltage characteristics may be a determining factor instead of the load-carrying characteristic. Also, it is possible that the insulating materials used in and around the resistor may have some influence on the rating. It is therefore customary to rate this type of unit on the

basis of the maximum load it can carry, or the maximum voltage that can be applied to it for a long time, without exceeding prescribed resistance changes. The prescribed changes generally accepted are 6 per cent for intermittent rated-load operation and 3 per cent for 625 per cent of rated load for 5 sec.

FIG. 11. Power derating curve for fixed composition resistors.

33. Power Derating Composition Resistors. When composition resistors are used at ambient temperatures exceeding 70°C, the permissible maximum power which may be applied is shown by the derating in Fig. 11.

34. Color Code. The resistance value and tolerance of both composition and wire-wound fixed low-power resistors are indicated by color bands. Such identification simplifies assembly of these units and helps in servicing. There is complete agreement among the various standards for color coding these resistors, and all manufacturers follow the standards.

Fixed low-power resistors with axial leads have their nominal resistance values indicated by bands of color around the body of the resistor, in accordance with the following system:

Band A indicates the first significant figure of the resistance value. If its width is the same as the other band it is a composition resistor. If its width is twice as wide as the other bands it is a wire-wound resistor.

FIG. 12

FIG. 13

FIG. 12. Fixed composition resistor with axial leads.
FIG. 13. Fixed wire-wound resistor (note band *A* is twice as wide as bands *B* and *C*).

Band B indicates the second significant figure of the resistance value.
Band C indicates the decimal multiplier.
Band D, if any, indicates the tolerance limits about the nominal resistance value.

The color code and examples are given below:

Color	Significant figure	Decimal multipliers: Power of 10	Decimal multipliers: Multiplying value	Tolerance, %
Black	0	10^0	1	
Brown	1	10^1	10	
Red	2	10^2	100	
Orange	3	10^3	1,000	
Yellow	4	10^4	10,000	
Green	5	10^5	100,000	
Blue	6	10^6	1,000,000	
Violet	7	10^7	10,000,000	
Gray	8	10^8	100,000,000	
White	9	10^9	1,000,000,000	
Gold	..	10^{-1}	0.1	± 5
Silver	..	10^{-2}	0.01	± 10
No color	..	....	....	± 20

Examples illustrating the standard are as follows:

Ohms	A	B	C
10	Brown 1	Black 0	Black, no cipher
220	Red 2	Red 2	Brown, one cipher
3,300	Orange 3	Orange 3	Red, two ciphers
4,700	Yellow 4	Violet 7	Red, two ciphers
68,000	Blue 6	Gray 8	Orange, three ciphers
150,000	Brown 1	Green 5	Yellow, four ciphers

35. Standard Resistance Values (Preferred Numbers). The industry has adopted the standard resistance values in accordance with a preferred-number series. These values have been adopted universally for fixed low-power resistors. This series of values, appropriately rounded to two significant figures, is defined for each tolerance band as follows:

Although EIA Standard GEN 102 and Military Standard MS-91374 are defined for somewhat narrower bands of resistance values, they are used by industry in the band of 0.1 ohm to 100,000 megohms.

1. For tolerance of ± 5 per cent, a series of values in which successive values are related by the multiplying factor $\sqrt[24]{10}$, *i.e.*, 1.10 approximately
2. For tolerance of ± 10 per cent, a series of values in which successive values are related by the multiplying factor $\sqrt[12]{10}$, *i.e.*, 1.21 approximately
3. For tolerance of ± 20 per cent, a series of values in which the successive values are related by the multiplying factor $\sqrt[6]{10}$, *i.e.*, 1.46 approximately

The table below gives the basic decade of standard preferred resistance values for each of the three standard tolerances.

Standard Preferred Resistance Values

Tolerances available, per cent					
5	10	20	5	10	20
1.0	1.0	1.0	3.3	3.3	3.3
1.1			3.6		
1.2	1.2		3.9	3.9	
1.3			4.3		
1.5	1.5	1.5	4.7	4.7	4.7
1.6			5.1		
1.8	1.8		5.6	5.6	
2.0			6.2		
2.2	2.2	2.2	6.8	6.8	6.8
2.4			7.5		
2.7	2.7		8.2	8.2	
3.0			9.1		

FIXED PRECISION RESISTORS

There are two classes of precision resistors, depending upon the nature of the conductor, namely, wire and film. Wire-wound precision resistors are usually used where low tolerance, 0.1 to 1 per cent, and high stability are required. Film types are used where the tolerances are of the order of ½ or 1 per cent and some sacrifice in stability

can be tolerated. For h-f applications, film types are almost always better because of their very small reactive components. Wire-wound precision resistors are considerably more expensive than precision film types, and thus their selection must usually be justified on a tolerance, stability, and economy basis.

36. Precision Wire-wound Resistors. These resistors are wound on forms, usually ceramic or plastic such as epoxy, having an even number of sections separated by thin barriers. The winding consists of enameled alloy resistance wire, multiple-layer-wound, adjacent sections being wound in opposite directions. The minimum wire size recommended is 0.001 in. over the enamel. The winding is usually continuous, although as many as four breaks, welded or brazed, are permitted in practice. Although adjacent sections are wound in opposite directions, the resistors are not noninductive. In audio circuits they do, however, help reduce inductive pickup. The wound resistor is then covered by an insulating coating to protect it against humidity. For less critical applications, the coating may be wax, varnish, or any other suitable material. For severe environmental conditions such as high humidity, resistors hermetically sealed in a ceramic container or encapsulated in epoxy resin or molded are preferred. Terminals used are wire leads or soldering lugs.

Precision resistors are available in a number of sizes and ranges as shown in Table 3. Although this table is not all-inclusive, it serves to indicate the general types that can be obtained. Table 4 lists the more common resistance alloys and their characteristics.

Table 3. Fixed Precision Wire-wound Resistors

Military type numbers	RB15	RB17	RB18	RB19
Power rating, watts*	½	1	1¼	2
Max voltage rating	300	700	1,000	2,000
Resistance min, ohms	0.1	0.1	0.1	0.1
Max resistance (for military use), megohms	0.4	1	3	6
Resistance max (for commercial use), megohms	1.5	4	7.5	20
Number of pie sections	2	4	6	8
Dimensions, in.:				
Length	7/16	31/32	1¼	2 3/32
OD	11/16	21/32	¾	1

Military-type epoxy encapsulated or molded, specification MIL-R-93A, commercial type, varnish coated.

* Power ratings based on 85°C ambient for military use, 65°C ambient for commercial use.

Table 4. Resistance-alloy Characteristics

Symbol	K	E	N	J
Alloy	Copper-nickel	High-resistivity nickel chrome	Nickel chrome iron	Nickel chrome
Trade names	Alloy 45, Advance, Copel, Cupron	Evanohm, Karma	Alloy C, Chromel C, Nichrome, Tophet C	Alloy A, Chromel A, Tophet A, Nichrome V
Resistance, ohms per cir mil ft at 25°C	294	800	675	650
Temp. coefficient of resistance, ppm/°C from 0–100°C	±22	±22	170	130
Thermal emf to CU, μv/°C from 0–100°C	−43	+3	+2	+5

Rating-Derating. It will be noted from the tables that there are two ratings, power and maximum voltage. The power rating is applicable to all resistance values up to that requiring the maximum voltage. The voltage rating applies to higher resistance values.

Derating curves are usually supplied by each manufacturer for their product. Typical curves for the epoxy encapsulated units are shown in Fig. 14.

37. Precision Voltmeter Multipliers. The more common construction of this type of precision resistor consists of a number of individual precision resistors interconnected and encased in a glazed ceramic or glass tube. Metal ferrule terminals are applied to the ends of the tube and are hermetically sealed by soldering to metal bands bonded to the ceramic. Variations in these construction techniques are sometimes employed to accomplish the same over-all result. These are available in three standard sizes with resistance ranges as follows:

Military type	MFA	MFB	MFC
Length, in.	9 25/32	5 9/32	2 15/16
Ferrule diam, in.	19/64	19/64	13/16
Min resistance, megohms	3.5	1.0	0.5
Max resistance, megohms	12.0	6.0	1.2

These resistors have a current rating of 1.0 ma maximum and the voltage rating in kilovolts is equal numerically to the resistance expressed in megohms. Standard tolerances are 0.5 and 1 per cent, but 0.1 per cent is available on special order. The military specification is JAN-R-29.

38. Film Resistors. *Deposited Carbon and Boron Carbon.* Deposited-carbon and boron-carbon precision resistors will be treated together, for these two types have much in common.

The deposited-carbon resistor is a film resistor, differing from the composition type in that the resistance film contains no dielectric material—only carbon. This film is deposited on a ceramic tube or rod by pyrolysis, *i.e.*, the cracking of a hydrocarbon gas at elevated temperatures in the proximity of the heated ceramic form. Under suitable conditions, crystalline carbon deposits on the ceramic to form an extremely hard dense film. Terminals may be wire leads, lugs, or ferrules. To obtain high resistance values, a helical groove is cut in the carbon coating, resulting in a long resistance path.

The process for making boron-carbon films is essentially the same as for deposited carbon, with the addition of a boron-containing gas being introduced during the cracking phase. This has the effect of reducing the T.C. over a limited range of values.

FIG. 14. Power derating curve for high ambient temperatures.

These resistors are provided with protective covers which may be multiple varnish coatings, plastic molding, ceramic casing, or a hermetically sealed glass case evacuated or filled with an inert gas. These insulated types usually increase in cost in the order given. The varnished type provides the least protection against mechanical abuse and humidity. For high reliability and military applications, the molded and encased units are preferred. Tables 5 and 6 give the basic information for these resistors.

Temperature Coefficient. The T.C. envelopes of deposited and boron-carbon resistors are shown in Figs. 15 and 16. The envelope shows the spread one can expect on T.C. on a lot-to-lot basis from different manufacturers.

High-frequency Characteristics. One advantage these resistors offer over precision wire-wound units is their frequency characteristic. Figure 17 shows the approximate frequency characteristic for coated ½-watt resistors. Coated resistors are somewhat better in frequency response than the equivalent molded units.

Table 5. Coated Deposited-carbon and Boron-carbon Specifications

MIL style	Power rating at 40°C ambient, watts	Max voltage rating, volts d-c	Body dimensions		Resistance values, deposited carbon	
			L length, in.	*D* diam, in.	Min, ohms	Max, megohms
RN32	1/8	250	9/32	5/64	50	1
RN10	1/4	300	15/32	3/32	10	1
RN20	1/2	350	9/16	5/32	10	2
RN25	1	500	15/16	19/64	10	5
RN30	2	750	2 1/16	19/64	30	10

Table 6. Molded Deposited-carbon and Boron-carbon Specifications, Military Specification MIL-R-10509B

MIL style	Power rating at 70°C ambient, watts	Max voltage rating, volts d-c	Body dimensions		Resistance values, ohms			
					Deposited carbon		Boron carbon	
			L length, in.	*D* diam, in.	Min	Max	Min	max
RN60	1/8	250	13/32	1/8	50	1 megohm	50	100,000
RN65	1/4	300	5/8	3/16	20	2 megohms	20	100,000
RN70	1/2	350	23/32	1/4	10	5 megohms	10	100,000
RN75	1	500	1 1/8	13/32	10	10 megohms	20	0.237 megohm
RN80	2	750	2 1/4	13/32	15	20 megohms	30	1 megohm

Standard tolerance is ±1 per cent. Can be supplied to ±0.5 per cent tolerance.

Temperature Rise. Figure 19 shows typical surface-temperature-rise data for the five standard sizes of molded deposited-carbon resistors.

Derating. Typical derating curves for coated and molded styles are shown in Fig. 18.

39. New Developments. "Carbon Alloy" Film. Research indicates that the inclusion of additives to the hydrocarbon gas during the cracking process may alter the properties of the film substantially by increasing permissible operating temperature, decreasing moisture sensitivity, and extending resistance coverage. So-called "car-

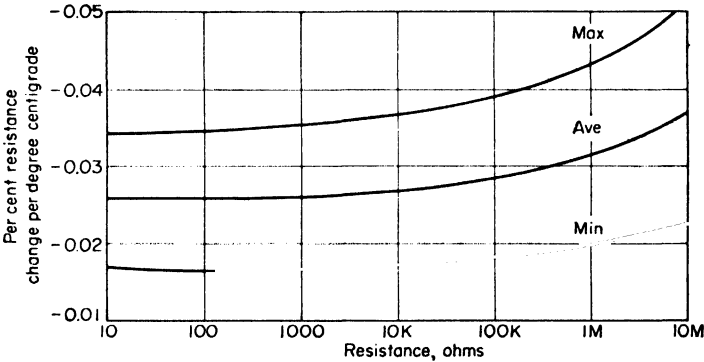



FIG. 15. Temperature-coefficient spread for deposited-carbon resistors.

bon alloy" film resistors operate satisfactorily at ambients of 200°C, and can be made as high as thousands of megohms.

40. Metal-film Resistors. There are several different varieties of metal-film resistors. In general, they offer the advantages of the deposited-carbon resistors. Some types have the additional advantages of extremely low and controllable T.C. over the entire range of values, and suitability for high-temperature operation.

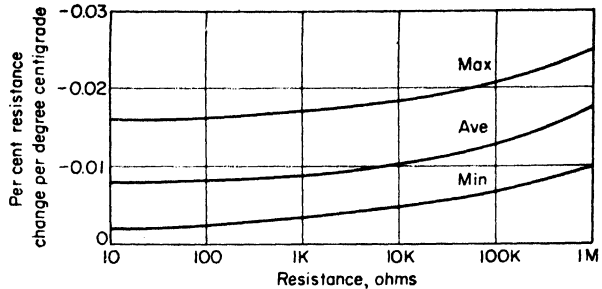



Fig. 16. Temperature-coefficient spread for boron-carbon resistors.

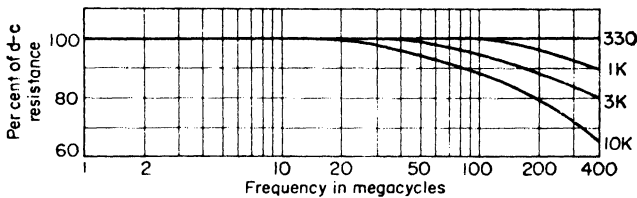



Fig. 17. Approximate frequency characteristics for ½-watt deposited-carbon resistor.

41. Evaporated Metal-film Resistors. These resistors are generally similar in physical appearance and dimensions to the deposited-carbon resistors. The resistance film is metallic and is deposited on a ceramic or glass form by evaporating a metal or alloy in vacuum. The metallic vapor condenses on the substrate to form a hard dense film. A helix is cut in the film, resulting in a long path and the desired resistance value. Terminations and protective covers are applied as with deposited-carbon resistors.

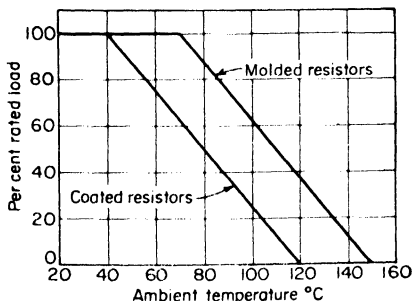



Fig. 18. Typical derating curves for coated and molded resistors.

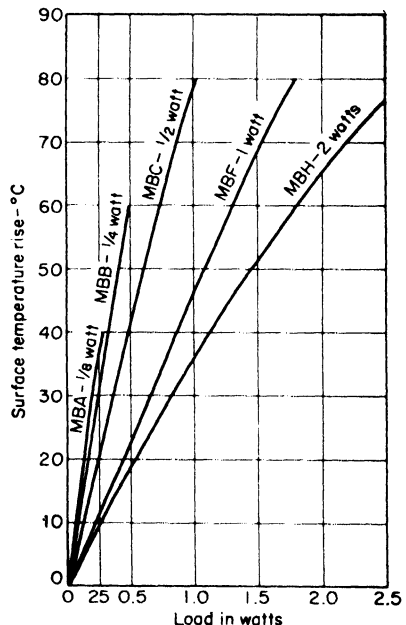



Fig. 19. Surface temperature rise vs. load, molded deposited-carbon resistors.

General Specifications. Physical dimensions and ranges of resistance values are generally the same as those shown for deposited-carbon resistors in Tables 5 and 6. The military specification applicable to these resistors is also the same, MIL-R-10509B. While this specification calls for operation at 70°C, some manufacturers offer full-load operation in ambients as high as 125°C. At least one manufacturer offers all five molded sizes called for in the foregoing specification, with temperature coefficients as low as 0 ± 25 parts per million per degree centigrade. Standard tolerance is ±1 per cent, but these resistors are also available in tolerances of ±0.5 per cent and ±0.25 per cent.

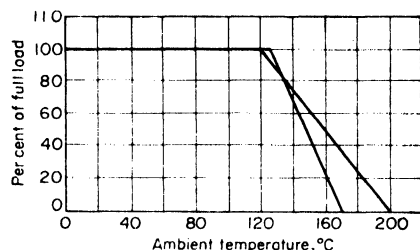



FIG. 20. Derating curve for metal film resistors, two different manufacturers.

Rating-Derating. Although the military rating is based on 70°C ambient, some manufacturers claim ratings at higher ambients. Figure 20 shows derating curves for two different manufacturers.

Hermetically Sealed Resistors. When these resistors are sealed in glass envelopes by conventional methods and the envelope is filled with helium, the rating is increased substantially. Operation at ambients as high as 150 to 200°C at full rating and 250°C at one-third rating has been accomplished. Such resistors have stability equal to precision wire-wound resistors.

Controlled Temperature Coefficient. The most interesting characteristic of the evaporated-metal-film resistor is shown in the graph of Fig. 21. This shows the T.C. range available for a wide array of resistance values for the various sizes of resistors. The line to the left of the crosshatched area represents the properties of nonspiraled resistors and shows how the T.C. varies with the thickness of the evaporated film. The

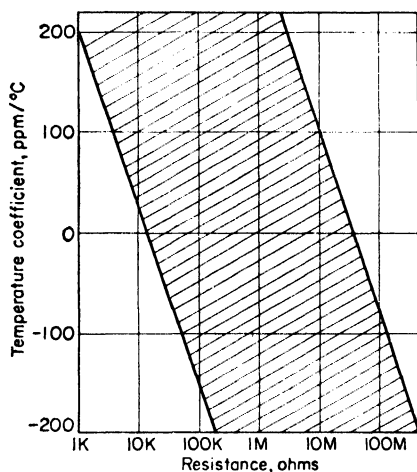



FIG. 21. Range of temperature coefficient available for resistors of various sizes.

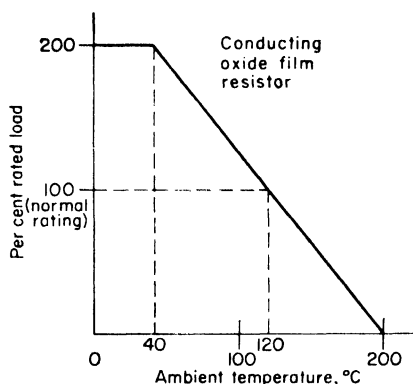



FIG. 22. Derating curve for conducting oxide-film resistors.

coefficient is positive for thick films; the magnitude decreases as film thickness decreases, crosses zero, and then turns negative for thin films. The point of zero T.C. is called the *crossover* point. This varies with the metal or alloy used. This phenomenon makes it possible to supply these resistors with controlled T.C. They can now be obtained on order to these T.C.s in parts per million per degree centigrade:

0 ± 50,	0 + 50,	0 − 50
0 ± 100,	0 + 100,	0 − 100
0 ± 150,		

42. Conducting Oxide-film Resistors. In this class of resistors, conducting oxides of tin and antimony are bonded at elevated temperatures to a glass rod or tube. The resistance element is then helixed to desired values. These are made, in general, to the same mechanical specifications as deposited-carbon and metal-film resistors.

Rating-Derating. For equivalent sizes and power rating of deposited-carbon and metal-film resistors, this type is rated by the manufacturer at 120°C ambient. The rating is doubled at 40°C ambient. This is shown in Fig. 22.

FIXED MEDIUM-POWER RESISTORS (WIRE-WOUND TO 20 WATTS)

These units dissipate between 2 and 4 watts per sq in. There are several different types in this class, the principal ones being:

43. Flat Wire-wound Medium Power. This type consists of wire wound on strips of fiber or laminated phenolic insulation to which lug terminals are attached at appropriate points. The strip is covered with phenolic either by molding or by other means. This assembly is then tightly enclosed in a sheet-metal punching with mounting holes, or a metal mounting strip is attached in intimate engagement with one side of the resistor, enabling the other side to be mounted flat against a metal chassis.

This design has many advantages. It is easy to mount. The metal enclosure, or mounting strip, and the chassis act as heat distributors and prevent excessive differences in temperature along the length of the unit. Use is made of the metal chassis and metal mounting to conduct heat away from the resistor. This makes possible higher power ratings for a given temperature rise. In effect this type of mounting approximately doubles the free-air rating of the resistor. For example, a resistor of this type, having a rating of 3.9 watts when suspended in free air, will have a rating of 7.4 watts when mounted on a steel plate 1 ft by 1 ft by 1/16 in.

They are used in the power range from 2 to 20 watts. Lengths vary from 2 to 6 in. Widths vary from 1/2 to 3/4 in. Terminals employed are generally soldering lugs.

44. Square-cross-section Medium-power Resistors. The resistance element consists of wire wound on a fiber-glass core assembled in a ceramic case of square cross

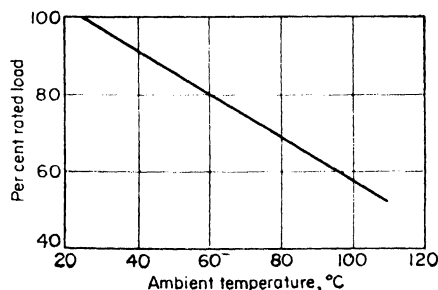



FIG. 23. Derating curve for resistors of square cross section of 5-, 7-, or 10-watt rating.

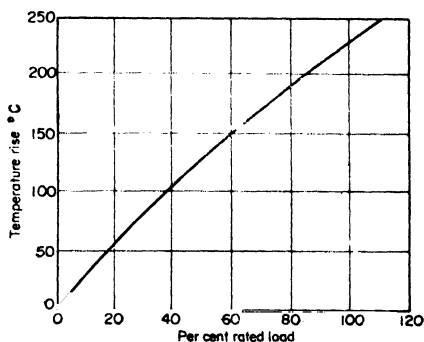



FIG. 24. Temperature rise vs. load for resistors of square cross section of 5-, 7-, or 10-watt rating.

section. The case is then filled with an inorganic cement which seals the resistor. The construction permits mounting against a flat metal panel for better heat conduction. It also permits magazining for automatic assembly to printed wired chassis. These resistors are supplied in 5 and 10 per cent tolerance. Typical dimensions and ratings are shown in the table at top of page 2-24. Figures 23 and 24 show typical derating and temperature rise for this class of resistor.

Power rating, watts	Dimensions		Resistance range, ohms
	Length, in.	Cross section, in.	
5	7/8	3/8 × 3/8	0.25– 3,300
7	1 3/8	3/8 × 3/8	0.5 – 9,000
10	1 7/8	3/8 × 3/8	5 –15,000

FIXED HIGH-POWER RESISTORS (10 TO 200 WATTS)

These resistors have the highest power ratings of the resistors normally used in electronic circuits. They dissipate 5 to 10 watts per sq in. of surface at 250°C rise from a 25°C ambient temperature when protected with inorganic materials. In the cases where higher powers are required, it is conventional to use several of these resistors in series.

45. High-power Wire-wound Resistors. These resistors are usually made by winding wire on ceramic forms and insulating and protecting them with inorganic and, for special applications, with organic materials. Many of the ceramic forms are circular in cross section but a "flat" style of elliptical cross section is also available for special or stack mounting. Most of these are protected with vitreous enamel or with cement. Some with circular cross section are hermetically sealed in tempered glass, some are made with many coats for protection against salt spray, and some are coated with an organic coating material for improved service in humid conditions but at reduced power ratings.

The terminals used on these resistors are wire leads, soldering lugs, or ferrules for mounting in fuse clips. Most manufacturers can supply L brackets with suitable insulating washers and through bolts for mounting these resistors. Some manufacturers also have Z brackets for mounting that do not require insulating washers and are less expensive but are more flexible.

These resistors are used for voltage dividers dissipating appreciable power, in grid circuits of high-power oscillators, and in a wide variety of circuits requiring high-power dissipation.

The most widely used power resistor is cylindrical in shape, and detailed specifications are given in military specification MIL-R-26B and industry specification EIA TR-132. The sizes, styles of terminals, power ratings, and ranges of resistance values are numerous. Typical information covering the ferrule-type resistor is given in Table 7.

Table 7. Standard Ferrule Types (JAN or MIL and EIA)

Type	Wattage	Max over-all length, in.	Body length, in.	Max shoulder diam, in.*	Fits form No. 1 std. clip fuse	Min ohms	Max ohms 0.0025 wire	Max ohms 0.00175 wire	Max ohms 0.0014 wire
RW10	140	11 5/8	10 7/16	1 13/32	60 amp 600 volts	3.0	71,000	210,000	350,000
RW11	116	9 13/16	8 5/8	1 13/32	60 amp 600 volts	2.5	63,000	175,000	310,000
RW12	86	7 5/8	6 7/16	1 13/32	60 amp 600 volts, 30 amp 600 volts	1.8	50,000	125,000	220,000
RW13	50	5 5/16	4 1/8	1 3/32	60 amp 250 volts, 30 amp 600 volts	1.2	25,000	60,000	100,000
RW14	40	4 5/8	3 7/8	1 3/32	60 amp 250 volts	0.9	16,000	48,000	85,000
RW15	20	3 1/8	1 15/16	27/32	30 amp 250 volts	0.5	6,300	16,000	28,000
RW16	14	2 9/16	1 3/8	27/32	30 amp 250 volts	0.25	4,000	9,000	16,000

* Body diameter is always less than shoulder diameter.

Most manufacturers offer many sizes that are not mentioned in these specifications. Examples of the range of sizes manufactured with lug-type terminals made by a manufacturer are given in Table 8. These resistors are available with taps and open sections on special orders.

Table 8. Standard Sizes

IRC type	Corresponding Jan or Mil types and rating		IRC wattage	Core dimension, in.			Min ohms	Max ohms 0.0025 wire	Max ohms 0.00175 wire	Max ohms 0.0014 wire
				Length	OD	ID				
1 A			5	1	5/16	3/16	0.10	530	1,550	3,000
1¾ A	RW29	8 watts	10	1¾	5/16	3/16	0.30	2,500	5,600	11,000
2 D			20	2	9/16	3/8	0.40	4,500	12,000	25,500
3½ E			40	3½	¾	½	0.85	13,500	40,000	77,000
4½ E			50	4½	¾	½	1.20	18,500	53,500	105,000
6½ E			80	6½	¾	½	2.00	28,500	82,500	160,000
6½ H			100	6½	1⅛	¾	1.80	41,500	120,000	236,000
8½ H			160	8½	1⅛	¾	2.50	57,000	167,000	322,000
10½ H	RW46, RW47	135, 145	200	10½	1⅛	¾	3.00	72,000	210,000	408,000

46. Special High-power Wire-wound Resistors. *Flat Type.* This is a wire-wound resistor using a rather flat elliptical ceramic with a hole of rectangular cross section through which the mounting bracket, made of a good heat-conducting metal such as copper or aluminum, is passed. The advantage of this type is that the heat is conducted from the unit, by means of the metal mounting bracket, to the chassis on which it is mounted. This results in a more uniform temperature distribution along the length of the resistor. The mounting hardware enables these units to be stacked one on top of another and so provides some space saving. However, the power rating of each resistor is reduced when used in this way because of the heat radiation from one unit of the stack to the other. The reduction in rating to be expected is as shown in Table 9.

Adjustable Resistors. In many applications a fixed resistance value is required, but the exact value is not known. The adjustable resistor fills this need. This is a standard cylindrical power resistor with a cement or vitreous-enamel coating. A portion of the form along its entire length, usually ¼ in. wide, is left uncovered, leaving the wire exposed. This forms a track on which an adjustable terminal, which makes pressure contact on the wire, may be moved until the proper value is obtained, after which the terminal is locked in position. Such resistors are made to tolerances of ±10 per cent only.

Noninductive Type. Power wire-wound resistors of all styles can be wound noninductively by using the Ayrton-Perry winding, consisting of two windings in opposite directions, with both windings connected in parallel. This type of winding has not only low inherent inductance but low capacitance as well if wound with bare wire. The crossover turns reduce the distributed capacitance because the adjacent turns have very little potential difference between them. Typical inductance specifications for this type of resistance are shown below.

Resistance Value, Ohms	Max Residual Series Inductance, μh, Measured at 1-3 Mc
1-99	0.5
100-999	1.0
1,000 or higher	2.0

Because of the double winding it is necessary to wind the noninductive resistor with very large pitch. This materially limits the maximum value of resistance obtainable on a given form, as shown by the following examples.

Size of winding form		Max resistance value, ohms	
Diam, in.	Length, in.	Inductively wound	Noninductively wound
5/16	1¾	12,000	2,000
9/16	3	37,000	2,000
3/4	4½	90,000	5,000
1⅛	6¼	200,000	5,000

47. Power Rating of Wire-wound Resistors. In view of the low temperature coefficient of the resistance wires generally employed in wire-wound resistors, the resistance change with loads normally encountered is small. The power rating is, therefore, primarily determined by the power that the resistor can dissipate continuously for an unlimited time without exceeding the permissible maximum temperature and without deterioration. This limits the permissible temperature rise, which varies for different classes of resistors, as measured from a specified ambient temperature, usually 25 or 40°C.

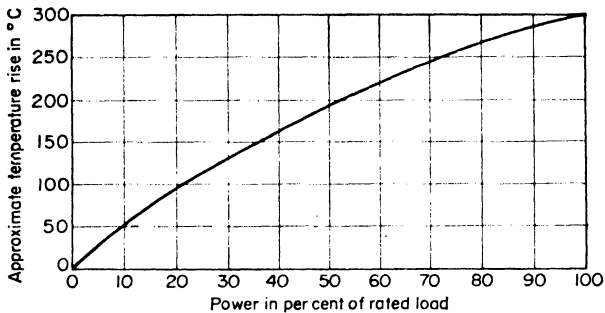



FIG. 25. Temperature rise for power wire-wound resistors in free air, inorganic coatings.

48. Temperature Rise of High-power Wire-wound Resistors. Although the temperature rise varies somewhat with different makes of resistors owing to differences in coating materials such as texture or color, Fig. 25 is typical and shows the temperature rise to be expected at various loadings of these resistors wound on ceramic forms, with vitreous-enamel and cement coverings. The 100 per cent rating is based on manufacturers' rating of 250°C rise in open air. Temperature is measured at the center of the outer surface of the resistor, the resistor being in free air. The factors influencing this rating are:

1. Heat-resistant properties of protective covering.
2. Heat-resistant properties of core.
3. Use of intermediate taps. Taps reduce effective winding space, resulting in less active cooling surface and thus reducing the nominal rating. The extent of reduction depends upon the length of the resistor, being smaller for long units than for short ones. On short units 2 in. long, the rating may be reduced by as much as 15 to 20 per cent; on units 6 in. long the reduction may be 3 to 5 per cent.
4. Ventilation. See Secs. 23 and 50.

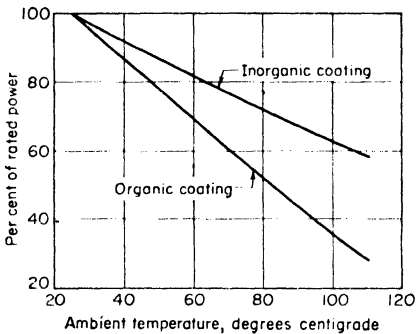



FIG. 26. Power wire-wound resistor derating curves.

49. Power Derating. In practice, the above considerations are taken into account by "derating" resistors. If a resistor is used at its nominal rating at an ambient temperature greater than its reference ambient, the maximum permissible temperature

will be exceeded. To avoid this, the r film on the inside of a glass tube. Water is fed
power low enough to compensate the y into the tube at a high velocity. It flows along
derating curves for each class of resisto ntact with the film and returns to its source.
operating load. This principle of dera vatts can be rated at 5 kw by such means. The
class of resistor can have such a derating cu
ture rise curve. Figure 26 shows typical ended. D-c operation destroys the resist-
tors having inorganic and organic coatings. r, etc., out of the water supply for they
specific type of unit, reference must be made
specification.

When "flat" resistors are stack-mounted the
exchange of heat with adjacent resistors by an a
general ambient temperature. These additional de

Table 9. Derating for Flat-type Resistors

Kind of stack mounting used	when n			
	2	3		
On vertical surface reduce rating by	18	30	35	
On horizontal surface reduce rating by	25	40	50	

FIXED HIGH-FREQUENCY RESISTORS

As commonly made, these resistors consist of ceramic tubes or rods with a resistive film deposited on the outer surface of the ceramic. This film may be a composition film described in Sec. 15 or any of the other film types described in Secs. 16 to 18. Conductive-film terminations are applied over the ends of the ceramic, to which leads, soldering lugs, or ferrule-type terminals are applied as required. For best h-f results the film should be continuous.

50. Composition-type H-f Resistors. Some of the characteristics are as follows:

1. The *temperature coefficient* varies with resistance value and varies from 0.005 per cent per °C for the minimum values indicated in Table 10 to 0.07 per cent per °C for the maximum values.

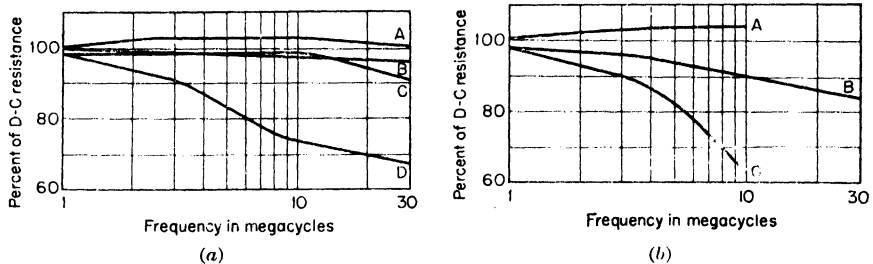
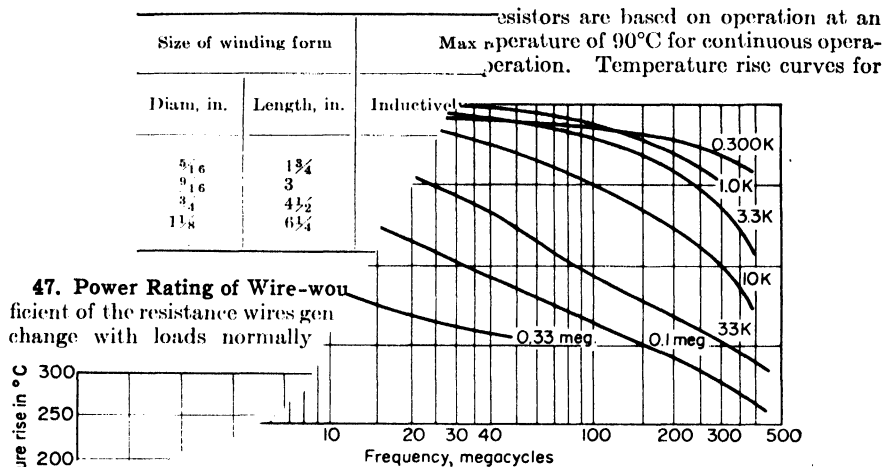



FIG. 27. Frequency characteristics of high-frequency resistors. (*a*) MPP resistor: *A*, 50 to 100 ohms; *B*, 1,000 ohms; *C*, 10,000 ohms; *D*, 0.1 megohm. (*b*) MPR resistor: *A*, 50 to 100 ohms; *B*, 1,000 ohms; *C*, 10,000 ohms.

2. The *voltage coefficient* also varies with resistance value and, between minimum and maximum resistance values, varies from approximately 0.002 to 0.03 per cent per volt per inch of length of resistance path. This limits the maximum voltage recommended.

3. Typical curves of the *frequency characteristics* are shown in Figs. 27 and 28 for different sizes and resistance values up to a frequency of 30 Mc for the higher-power units and 400 Mc for the low-power units.

...esistors are based on operation at an ...perature of 90°C for continuous opera... ...peration. Temperature rise curves for

47. Power Rating of Wire-wou... ...ficient of the resistance wires gen... change with loads normally ...

Fig. 28. Typical frequency characteristics, 2-watt high-frequency resistors.

this type resistor are shown in Fig. 29. If operated at higher ambients, it is necessary to derate in accordance with the derating curve in Fig. 30.

Increasing Power Ratings. This type of resistor, when operated in free air, is restricted to medium power ratings because of the temperature limitations of the materials used in the resistive film. Means for increasing the power rating are:

1. *Forced Air-draft Cooling.* The use of an air stream from a fan or blower can increase the rating from two to three times.

2. *Immersion in Oil.* This also has the effect of increasing the rating by a factor of 2 to 3.

Fig. 29. Type MP resistor; load-temperature rise in free air.

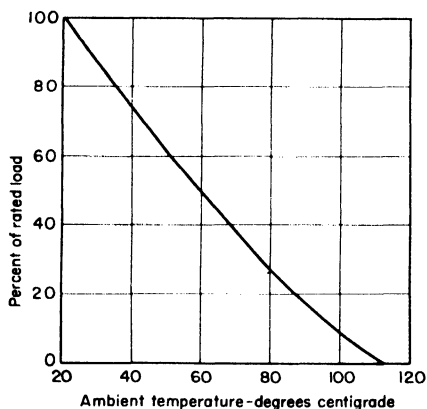



Fig. 30. Derating curve, type MP resistor.

3. *Water Cooling.* Extremely large increases in rating from 100 to 250 times the air rating can be obtained by appropriate methods of water cooling. This type of water-

cooled resistor is made with the resistor film on the inside of a glass tube. Water is fed through appropriate fittings tangentially into the tube at a high velocity. It flows along the resistor film in intimate thermal contact with the film and returns to its source. Resistors having a free air rating of 20 watts can be rated at 5 kw by such means. The rating depends upon the water flow.

Operation below 0.5 Mc is not recommended. D-c operation destroys the resistance film. Keep sand, grit, lumps of solder, etc., out of the water supply for they may cause early failure of the resistor. The power rating varies according to water flow, as shown in Fig. 31.

51. Carbon-alloy and Metallic-film H-f Resistors. Resistors employing films such as described in Secs. 38 and 40 have recently been introduced for h-f applications. The general construction is the same as for the composition type. The principal advantage of these films is their ability to operate at much higher temperatures than the composition films. The maximum operating temperatures of these types may reach as high as 225 to 250°C. In general, the power ratings of these resistors of a given size are approximately three to four times as great as those for composition resistors. As a result, these may be used for high-power h-f applications in place of power wire-wound resistors where their reactive component is too large. For metal-film resistors the T.C. is less than 0.03 per cent per °C and voltage coefficient is less than 0.001 per cent per volt. Table 10 gives a few typical sizes with pertinent information.

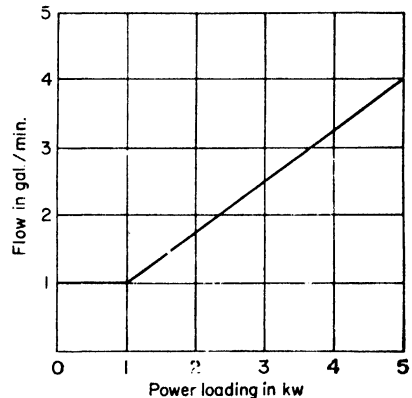



Fig. 31. Power rating of water-cooled resistors vs. flow of cooling water.

Table 10. High-frequency Resistors with Lug Terminals or Colloidal Silver Terminals

Type	Power rating, watts	Peak voltage rating	Min available resistance, ohms		Max available resistance, megohms ± 5%, ± 10%, ± 20%	Dimensions, in.			
			± 20% or ± 10%	± 5% tolerance		Body length	Resistor body diam	Ceramic ID	Width of silver bands No. 15 terminal
MPF	2	500	20	100	1.5	1¾ ± 1/32	5/16 ± 1/32	3/16 ± 1/32	3/8 ± 1/32
MPG	4	500	20	100	1.5	2 ± 1/32	9/16 ± 1/32	3/8 ± 1/32	9/16 ± 1/16
MPJ	5	1,000	35	200	3.0	3 ± 1/32	9/16 ± 1/32	3/8 ± 1/32	9/16 ± 1/16
MPP	10	1,750	35	150	2.5	4½ ± 1/16	3/4 ± 1/32	1/2 ± 1/32	9/16 ± 1/16
MPA	20	2,750	35	250	4.0	6½ ± 1/16	1⅛ ± 1/32	3/4 ± 1/32	9/16 ± 1/16
MPE	25	3,800	40	300	4.5	8 5/16 ± 1/16	1⅛ ± 1/32	3/4 ± 1/32	9/16 ± 1/16
MPO	30	4,750	50	400	5.0	10½ ± 1/16	1⅛ ± 1/32	3/4 ± 1/32	9/16 ± 1/16
MPR	90	8,500	60	500	6.0	18½ ± 1/8	2 ± 3/64	1 9/16 ± 1/32	7/8 ± 1/16

HIGH-VOLTAGE RESISTORS

The increasing use of high-voltage sources, as in radar, television, and X-ray equipment, introduced a requirement for resistors capable of handling high voltages at moderate power. Also, these resistors must be made to extremely high values, hundreds to thousands of megohms, so that the load on the high-voltage source may be low.

52. High-voltage Resistor Characteristics. High-voltage resistors are made by applying a resistance film on a ceramic tube as described in the previous section on

h-f resistors, except that the resistance path is a continuous spiral line. In this way a long resistance path can be obtained on a very short tube. For example, the resistance path on a tube 3 in. long and 9/16 in. in diameter can be made to have a linear length of 33 in. On the basis of a 300-volt-per-in. rating, such a resistor could be rated for 10,000 volts. The long resistance path also makes it possible to obtain very high resistance values more readily. For example, on the basis of 600 megohms per lineal in. of resistance path, which is readily obtainable with composition films, it is possible to obtain 20,000 megohms in this same resistor.

Various kinds of resistive films are used such as composition films, deposited-carbon films, conducting-oxide films. The most widely used thus far have been the composition films. These resistors are made in many sizes and ratings. Table 11 gives information on typical sizes of composition-film resistors as supplied by one manufacturer. The use of conducting-oxide films and carbon-alloy films makes it possible to operate these resistors at temperatures as high as 225 to 250°C.

Table 11. High-voltage Resistors

Types	Power rating, watts	Peak voltage rating	Dimensions, in.		Resistance, ohms		Resistance path length, in.
			Length	Diam	Min	Max	
MVX-1	1	3,500	1 1/16	1/4	2,500	2,500	5
MVF	2	5,000	1 3/4	5/16	25,000	15,000	10
MVJ	5	10,000	3	9/16	0.2 megohm	20,000	33
MVO	30	50,000	10 1/2	1 1/8	2.0 megohms	20,000	330
MVR	90	100,000	18 1/2	2	7.0 megohms	20,000	1,050
MVM-6	...	30,000	6	5/16	75,000	1,500	77

The maximum allowable operating temperature is 100°C.

Voltage Coefficient. This varies somewhat with the manufacturer and may be important in some applications. It is a function of the voltage per unit length of the resistance path and of the resistance per unit length of the path.

The voltage coefficient of MV resistors shown in Table 11 varies from approximately −0.002/RPL to −0.03/RPL per cent per volt, depending upon resistance (RPL = resistance path length). The active length of the resistance path is given in Table 11 for calculation of approximate voltage coefficient.

Example: To determine a voltage coefficient for a low-range MVF resistor,

$$\frac{-0.002}{10} = -0.0002\,\%/\text{volt}$$

Temperature Coefficient. This is approximately −0.005 per cent per °C for low resistance values to −0.1 per cent per °C for high resistance values over the temperature range of −55 to +105°C.

VARIABLE RESISTORS

53. Variable Composition Resistors. In numerous applications high-valued variable resistors are required; *e.g.*, for controlling the sensitivity of a receiver by varying the bias on the r-f tubes a variable resistor up to 50,000 ohms maximum is commonly employed. For adjusting the audio signal level in a-v-c radio sets a variable resistor up to 2.5 megohms is not uncommon. From the point of view of cost, wire-wound resistors of this magnitude are prohibitive. Furthermore, it is frequently desirable to have a nonuniform rate of change of resistance with respect to angular rotation, which is very difficult to secure with wire-wound resistors. Therefore, carbon or graphitic types of variable resistors made to meet these requirements at reasonable cost are widely used.

There are two classes of this type of resistor:

1. *Film Type.* Such resistors generally consist of a resistive solution applied to some flat form, such as paper, plastic, or ceramic, and baked on. The resistor solution consists generally of carbon or graphite, or both, mixed in a resinous binder (varnish) or lacquer. This solution may be applied to the insulating base by spraying, rolling, printing, or other means of depositing a liquid. A rotating slider or some other form of contact travels over this resistive element and produces a continuous variation of resistance. Since the resistor is essentially painted on the form, its geometrical form may be varied by design. Also different concentrations of the resistor ink or paint may be employed at different positions of the resistor element. By the use of these two expedients, the resistor may be designed to give any resistance-rotation law desired.
2. *Solid Type.* The resistor element consists of carbon and graphite mixed with a resinous molding powder and molded in solid form to an insulating base. A rotating slider or other form of contact travels over this element and produces a continuous variation in resistance. By varying the ratio of conductor to binder along different positions of the resistor track, the resistor element may be made to follow different resistance-rotation laws.

In both types the resistance element is mounted on or embedded in an insulating base, and this assembly is enclosed in a metallic case. A line power switch is frequently assembled to the cover, which is operated by the shaft. Both types are made in resistance values up to 10 megohms. Standard tolerance for this type unit is 20 per cent, although they are also made to 10 per cent.

The mechanical characteristics are as follows:

1. *Size.* These controls are made in sizes varying from ½ to 1¼ in. in diameter. Standard sizes are ½, ¾, 15/16, 1⅛, and 1¼ in.
2. *Mounting.* Single-hole mounting is standard, using a ⅜ in. by 32 threads per inch as the standard mounting bushing for the larger controls.
3. *Rotation.* Rotation is by means of a standard ¼-in.-diameter shaft for the larger controls to which a knob is attached.
4. The *terminals* are solder lugs.
5. *Taps.* Taps or terminals may be brought out at intermediate points along the resistance element for any purpose, *e.g.*, for acoustic compensation at low levels. The standard positions for bringing such taps out are 35 and 65 per cent of effective electrical rotation.
6. *Multiple Controls.* Standard controls are made so that a number of circuits may be controlled simultaneously. Two or more single controls are mounted on one shaft and bushing assembly, so that rotation of the shaft will vary all the controls simultaneously.
7. *Concentric Dual Controls.* There are equipments, such as television receivers, in which a large number of circuits must be controlled independently. The use of individual single controls would require considerable space. So-called "concentric dual controls" are made to avoid this loss of space. These consist of a two-section control with one mounting bushing, using concentric shafts, the hollow shaft controlling the front section, the solid shaft controlling the rear section. Two circuits can be controlled independently, and the panel mounting space is that required for only one control.

The electrical characteristics are as follows:

1. *Voltage Coefficient.* Since the composition of the resistor element is the same as for fixed-composition resistors, the voltage characteristic is also the same.
2. *Temperature Characteristics.* The same considerations apply here as to fixed composition resistors. Figure 32 shows typical characteristics for a volume-control resistance element on phenolic sheet, for different resistance values.
3. *Noise.* Standard quantitative requirements and methods of measurement have not yet been established for this characteristic. The characteristic of importance is the noise generated as a result of the movement of the contactor across the element. This is a function of a number of factors:

a. Surface of resistor element. A smooth flat surface produces less noise than a rough irregular one.

b. Nature of solid contactor. The points or area of the slider at which contact is made to the resistance element should not change or shift as the slider is rotated, if minimum

noise is to be obtained. Such shifts produce irregular resistance changes and current distribution which produce noise. This imposes severe requirements on the flatness of both resistance element and contactor.

c. Taper or resistance-rotation curve. Smoothness and gradualness of a taper curve are important for minimum noise. Bad breaks, irregularities, and sharp changes in slope in the curve cause corresponding changes in resistance which produce noise.

d. Presence of d.c. in the resistor circuit aggravates the production of noise. This explains why diode load and sensitivity controls in cathode circuits used as rheostats are more noisy than audio-level controls.

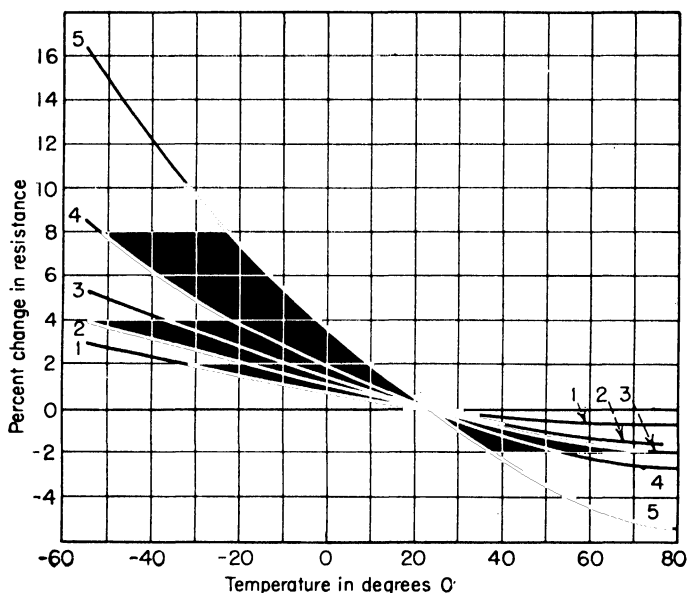



FIG. 32. Temperature-resistance characteristics; printed resistance on phenolic sheet. Curve 1, 1,000 ohms; curve 2, 10,000 ohms; curve 3, 0.100 megohm; curve 4, 1.00 megohm; curve 5, 10.0 megohms.

4. *Power Rating.* The power rating of the film-type variable resistors is ¼ watt for the 15/16-in.-diameter size, ⅓ watt for the 1⅛-in. size, and ½ watt for the 1¼-in. size.

The solid type of variable composition resistor has a power rating of 2 watts for the 1⅛-in. size, and ½ watt for the ½-in. size. Rating based on 70°C ambient.

54. Uses for Variable Carbon Resistors. Within their power limitation these resistors may be used wherever a continuously variable resistor is required. They may be used as either potentiometers or rheostats. They find their widest use as volume controls and tone controls in radio receivers. Some of their specific uses are here listed, and some of the basic circuits illustrating these uses are shown in Fig. 33.

1. Sensitivity control for radio receivers, by varying control-grid or screen-grid potentials of r-f tubes (Fig. 33*a*).
2. Antenna control for varying r-f input to antenna tube (Fig. 33*b*).
3. Sensitivity and antenna input control, combination of Figs. 33*a* and 33*b* (Fig. 33*c*).
4. Audio-level control (Fig. 33*d*).
5. Combination load-resistor and audio-level control in diode rectifier circuit (Fig. 33*e*).
6. Tapped volume control for acoustic compensation at low levels. Resistance-capacitance circuits are shunted across one or more taps to produce varying degrees of a-f compensation at different levels (Fig. 33*f*).

7. Gain controls and faders for phonograph and a-f amplifiers.
8. Tone control in a-f amplifiers for varying a-f frequency characteristics.
9. High-frequency variable resistor when nonreactive feature is essential, as in signal generator attenuators.
10. Television controls, such as brightness, contrast, focusing.

55. Tapers. The circuit considerations involved in these applications are discussed elsewhere in this handbook, particularly in Chap. 19, Receivers. However, each of these applications calls for a resistance curve, or "taper" as it is termed, which is most

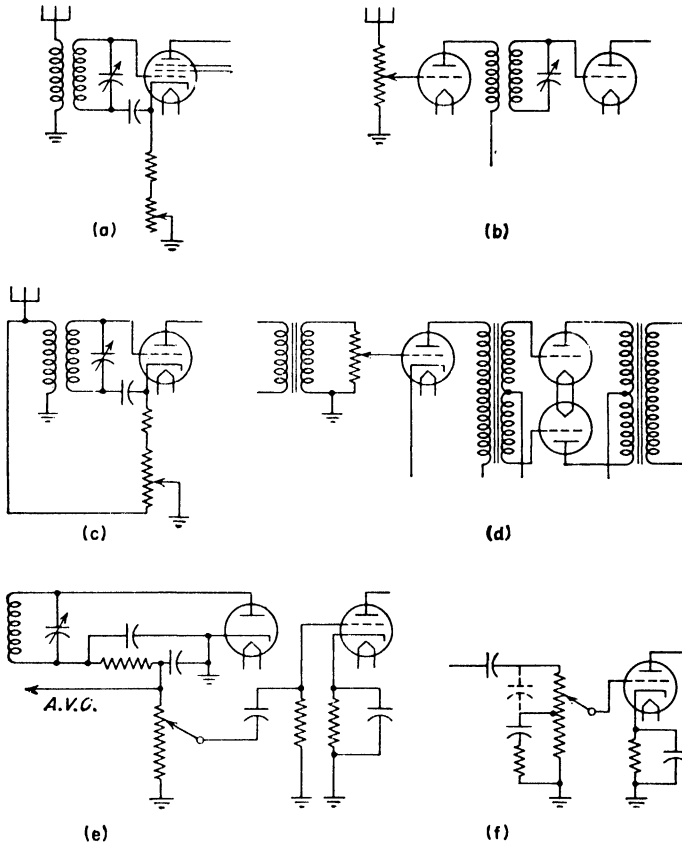



FIG. 33. Typical uses of variable resistors.

suitable for it. This taper defines the law of resistance change vs. angular rotation of the variable arm. Some widely used curves are given in Fig. 34.

A suitable specification defining the taper should include:

1. Curve showing resistance variation against active angular rotation of the contactor. Where a switch is incorporated in the variable resistor, the angle taken up for operation of the switch is considered inactive. Curve should indicate whether resistance increases with clockwise or counterclockwise rotation.
2. Resistance at extreme counterclockwise end between variable arm and left terminal; this is generally called "left terminal minimum" and is specified as "less than so many ohms."

3. Resistance at extreme clockwise end between contactor and right terminal; this is generally called "right terminal minimum" and is specified as "less than so many ohms."

4. When a tap is specified, the angular location and resistance of the tap should be given. The resistance between the tap terminal and the variable arm, when located at the tap, is sometimes specified.

56. Choice of Volume-control-resistance Curve. In an audio amplifier in which the maximum output is 40 db above the minimum output, the volume control should be so made that each $\frac{1}{40}$ of the rotation should correspond to an attenuation of 1 db. If the volume control has a total attenuation of 80 db, more than is necessary on this particular amplifier, each $\frac{1}{40}$ of the rotation will correspond to 2 db attenuation since only half of the total rotation can be used. In the second case the control should be more critical than in the first case.

FIG. 34. Typical taper curves for variable resistors.

In a radio receiver the design of the volume control differs widely depending upon whether the receiver has a.v.c. or not. If not, the entire voltage gain of the receiver must be under control, perhaps 120 db. The tendency for the volume control to become noisy or to be difficult to adjust without producing violent jumps of volume change increases with the total gain that must be controlled.

The fact that a-v-c systems cannot deliver a uniform voltage to the audio detector because of the wide variations of input voltage (ranging from a microvolt to several volts) makes necessary a different shape of attenuation curve than would be used on an audio amplifier used by itself. A type of curve useful in the a-v-c receiver is shown. Here, approximately uniform attenuation of 40 db is secured in 80 per cent rotation from the maximum volume. This is the range most often used. The departure from linearity in the first 15 per cent of rotation is to keep the resistance gradient within limits representing low noise (Fig. 35).

Between 80 and 100 per cent rotation, the curve changes rapidly to provide a total attenuation of 80 db. Rapid attenuation in this region is accomplished without noise because the resistance change per decibel is small. Such a curve is much more satisfactory than a straight logarithmic line (note the 80-db curve). Units of this type are simpler to build. A tapered resistance curve such that equal increments in rotation produce equal increments in attenuation (a straight line when plotted against the logarithm of the resistance) requires that a change of 300,000 ohms take place in the first 10 per cent, 120,000 ohms in the second 10 per cent, and so on till the last 10 per cent rotation produces a change of only 75 ohms. This is true of a 500,000-ohm control with a total attenuation of 80 db.

Wear Characteristics. Variable carbon resistors necessarily have the same general electrical characteristics as fixed carbon resistors. In addition, owing to the motion of the slider on the resistance element, there is a certain amount of wear on the resistance element. This produces a change in resistance value and noise. Factors influencing these changes are as follows:

1. Hardness of resistance element which determines ability to withstand abrasion.
2. Pressure of moving contact on resistance element.
3. Smoothness of moving contact surface.

57. Variable Wire-wound Resistors. 1. *Low Power.* These are usually of the continuously variable type made by winding resistance wire on an insulating form of bakelite, fiber, or similar material, which may be arcuate in shape or a flat strip formed into an arc after winding. The form is then placed in a protective container. A metallic sliding arm is arranged to travel over the winding, thus making contact with each turn as it is rotated. The choice of wire and size is determined by the resistance and space requirements.

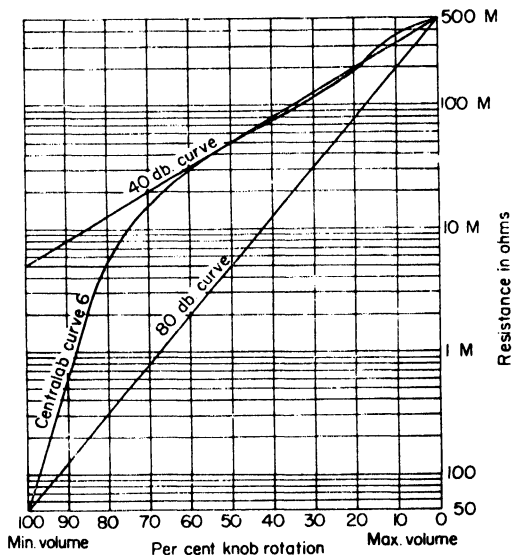
In general, wire-wound continuously variable resistors are wound so that the resistance changes uniformly with the motion of the sliding contact. For certain uses, it is desirable that the resistance change be nonuniform. In this case, the form on which the wire is wound is sometimes tapered so that the resistance per degree of rotation is not constant. Other methods of tapering employed are winding with variable pitch, winding sections of the control with different sizes of wire, and copperplating portions of the winding.

FIG. 35. Advantage of special taper for volume control.

The resistors are made in sizes varying from approximately $1\frac{1}{4}$ in. in diameter and $\frac{1}{2}$ in. deep to 5 in. in diameter and $2\frac{1}{2}$ in. deep. The maximum resistance varies from 10,000 ohms for the smallest size to 500,000 ohms for the largest. The standard power ratings vary from 2 watts for the smallest size to 25 watts for the largest. Usual standard ratings are 2, 3, 7, 8, 12, and 25. The power rating is based on a temperature rise of 60°C, when mounted on a steel plate $\frac{1}{16}$ in. thick and 4 to 6 in. square. This rating is based on a resistor wound linearly.

2. *High Power.* The general type of high-power variable resistors is the ceramic type. Resistance wire is wound on a toroidal ceramic core, and the winding coated with vitreous enamel and bonded to a ceramic base or otherwise mounted on materials having high temperature resistance. A sliding contactor, which may be metallic, or a graphite-metal brush is arranged to rotate around an open track on the resistor winding.

These resistors are made in sizes varying from $1\frac{1}{2}$ to 12 in. in diameter. They are made in power ratings from 10 to as high as 1,000 watts, the usual standard ratings being 25, 50, 75, 100, 150, 225, 300, 500, 750, and 1,000 watts. Standard tolerance is ±10 per cent. The maximum resistance values that can be wound are very high, based on the size of these rheostats. Since these units are intended for relatively low voltage systems in which heavy currents and high power are controlled, the maximum resistance values are generally limited to 5,000 to 10,000 ohms.

Power Rating. Power rating is the power dissipated in the total resistance of the rheostat. Linear-wound rheostats of this type having ratings of 100 watts and less are rated on the following basis: When mounted in the center of a 0.063-in.-thick steel panel, 12 in. square and vertically suspended in still air at an ambient of 40°C, the temperature rise, measured at the center of the open track, at rated power, shall not exceed 300°C. For

rheostats above this rating, under similar conditions, the temperature rise shall not exceed 350°C. Enclosed rheostats are rated at 50 per cent of this unenclosed rating.

Design Consideration. Some of the factors to be considered in design are as follows:

1. Contact between slider and resistor element should be positive.
2. Winding should not become loose on the form.
3. Sliding contact should not wear away resistance wire.
4. Resistance change per turn should be as small as possible.
5. Slider material should be such that it will not oxidize, wear excessively, or pit.

PRECISION POTENTIOMETERS

These are variable resistors designed to give a precise relationship between the displacement (linear or angular) of the slider and the resistance or voltage at the point of contact of the slider. This relationship may be linear or functional (nonlinear). These find their greatest applications in systems requiring accurate positioning such as servomechanisms, fire control, radar, computers, and automatic control devices.

58. Potentiometer Types. There are two types:

1. *Rectilinear.* The resistance element is straight and the slider displacement is linear.
2. *Rotary.* The resistance element is circular and the slider displacement is angular. The rotary type may be a single-turn potentiometer (360 deg or less rotation) or multiple-turn. Multiple-turn potentiometers are made with rotations as high as 5,400 deg or more. The discussion of any type is applicable generally to all types.

59. Potentiometer Construction. The basic elements of the precision potentiometer are the same as for any variable resistor, namely, a housing in which are assembled a resistance element and a movable arm or shaft carrying a contact which slides along the resistance element to vary the resistance. The difference resides in the precision, accuracy, and care with which every part and the assembly are made, and in the requirements which have to be met.

The resistance element is the heart of the device and is assembled in a plastic or metal case. Resistance wire is wound on an insulating form which generally is either a flat phenolic strip of rectangular cross section, or Formvar-insulated copper rod of uniform cross section. Both types may be used straight or shaped into a circle for a single-turn potentiometer. The copper rod winding may also be shaped into an accurately spaced helix to provide a multiturn potentiometer. When flat strip is used for the winding, it may be tapered or shaped to a contour to provide a nonlinear resistance change.

For a film-type resistance element the insulating base is generally phenolic for composition films, and either glass or ceramic for deposited-carbon or metal films.

The shaft or movable element carries a slider or contact which is generally a composition of noncorrosive precious metals designed to give minimum wear and noise. The shaft rotates in a sleeve or ball-bearing assembled to the case. Ball bearings are used where extremely low torque is a requirement. The case is enclosed to be dust-free. The design is such that a number of potentiometers may be assembled in tandem operated by one shaft.

Dimensions of the rotary type vary from ½ to 3 in. diameter and ½ to 3 in. long. Power ratings vary from ½ to 10 watts. The resistance varies from a few ohms to approximately 500,000 ohms, depending upon the size of the potentiometer.

60. Potentiometer Performance. The unique factors determining the electrical performance are linearity, conformance, resolution, and noise.

1. *Linearity.* This is a measure of the ability of a linear potentiometer to produce an output (voltage or resistance) directly proportional to the slider displacement (angular rotation of shaft). There are many types and definitions of linearity. Figure 36 illustrates a typical linear potentiometer in which the ideal or desired curve of resistance change with rotation is shown as a straight line. The actual curve might appear as shown. The linearity of the potentiometer is defined as the maximum per

cent deviation from the desired curve. When the curves are plotted in per cent resistance vs. per cent rotation, the per cent linearity is the difference between the two curves at the point of maximum deviation. Standard specifications for linearity usually are of the order of 0.5 to 1 per cent, but on special orders may be made as low as 0.1 per cent.

Good linearity is dependent on factors such as:
a. Uniform cross section of winding forms
b. Uniform specific resistivity of the wire
c. Uniformity of the winding pitch and its maintenance
d. Good mechanical construction such as concentricity of shaft, bearing and winding, absence of backlash

Improvement in linearity can be obtained by the use of trimming resistors at one or both ends of the potentiometer and by padding resistors at tapped points on the potentiometer.

2. *Conformance.* This term applies to functional or nonlinear potentiometers and has the same significance that linearity has to linear potentiometers. It is a measure of the degree to which the actual output curve deviates from the desired curve.

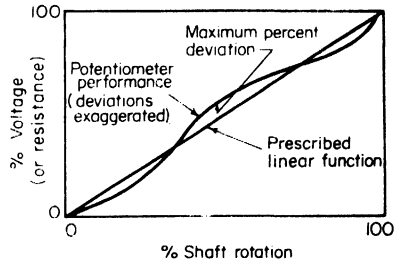
Nonlinear potentiometers are made by shaping the winding form to correspond to the desired function. In this case, lack of conformance is frequently corrected by connecting external or internal padding resistors either at both ends and/or at tapped points along the winding to distort the curve shape until the desired curve is obtained.

FIG. 36. Linearity of precision potentiometers.

3. *Resolution.* This is a measure of the precision to which the output of the potentiometer can be set. It is measured by the minimum resistance change that can be obtained by motion of the contactor expressed as a percentage of the total potentiometer resistance. The smaller the resistance change the greater the resolution. In a wire-wound potentiometer of a given resistance, therefore, the greater the number of turns the better the resolution. It is for this reason that multiturn potentiometers are so widely used, since the number of turns of wire for a given resistance increases with the number of revolutions of the shaft. However, no matter how many turns of wire are used, the minimum resistance change obtainable is the resistance of one turn.

The ideal condition, therefore, for precision setting occurs when there is a continuous stepless change in resistance. This is most nearly obtained with film-type and slide-wire resistors. In both cases the resistance change is continuous, the resistance element having the equivalent of an infinite number of turns, producing an *infinite-resolution potentiometer.* The slide-wire type, however, is severely limited in the maximum resistance value that can be obtained. This disadvantage can be partially offset by a multiturn slide-wire potentiometer. The importance of a very low resolution is evident in the case of servos where "hunting" may take place in stepwise control, the servo trying to find a resistance value or voltage which does not exist. Further, the importance of low resolution is also evident in the last important requirement of precision potentiometers, namely:

4. *Noise.* Whatever the causes, noise is the equivalent of unwanted resistance changes or spurious voltages which transmit false information. Some of the causes of noise are:

a. Resolution factor. A potentiometer with low resolution, *i.e.*, few turns or large resistance change per step, results in abrupt changes in voltage as the contact travels from turn to turn. A high resolution or large number of turns are necessary for minimum noise.

b. High contact resistance between contacts and resistance element. This may be caused by insufficient pressure of contactor to wire. Too great contact may cause too much wear. A compromise must be made.

c. The presence of dust, dirt, wear products, or other foreign matter on the resistance element. Extreme care must be exercised in manufacture to avoid this.

d. Oxidation of the resistance element of the contactor.

e. The presence of emfs in the potentiometer as a result of thermal emfs, or polarization emfs developed by dissimilar metals or impurities in the metals. This effect is extremely small.

61. Potentiometer Specifications. There are a great many variations in this class of potentiometers. Complete standardization has not yet been effected. Table 12 below is an illustration of the specifications covering a small segment.

Table 12. Typical Specifications of Precision Potentiometers

Multiturn Potentiometers

No. of turns	Case diam, in.	Case length, in.	Resistance range, ohms		Resistance tolerance, %		Linearity tolerance, %		Power rating, watts	Terminal type	Starting torque, oz-in. max	Shaft diam, in.	Weight, oz
			Low	High	Std. ±	Best ±	Std. ±	Best ±					
10	1½	2	150	500K	5	0.5	0.5	0.1	6	Turret	2	¼	3.1
5	1	1⅜	50	175K	5	0.5	0.5	0.1	3	Turret or flex	2	¼	1.8
10	¾	1¹¹⁄₁₆	50	250K	5	0.5	0.5	0.1	3	Turret or flex	2	¼	1.4
10	½	1¹⁵⁄₃₂	50	50K	10	1	1	0.25	1.5	Turret	2	¼	1
Single-turn Potentiometers													
1	1	9⁄16	10	35K	5	0.5	0.5	0.25	1½	Turret	1	¼	1
1	¾	⅜	50	25K	10	1	1.0	0.5	1	Turret	1	¼	1
1	½	²¹⁄₆₄	5	5K	10	3	5	2.5	½	Turret	1	3⁄16	¼
1	3	1.086	1K	600K	5	1	0.5	0.1	6	Turret	0.25	¼	6.25
1	2	1.086	500	300K	5	1	0.5	0.15	4.5	Turret	0.22	¼	3
1	1¾	0.863	500	200K	5	1.5	0.5	0.15	3	Turret	0.18	¼	3
1	1⁷⁄₁₆	0.863	100	150K	5	1.5	0.5	0.15	2	Turret	0.15	¼	2
1	1¹⁄₁₆	0.765	100	100K	5	2	0.5	0.2	1.4	Turret	0.12	⅛	0.75
1	⅞	0.765	100	80K	5	2	0.5	0.25	1	Turret	0.10	⅛	0.5

VARISTORS AND THERMISTORS

62. Varistors. Varistors is the coined name applied to a class of nonohmic conductors, whose resistance, though not manually adjustable, varies over wide ranges with voltages and/or temperature. These are made of a class of materials called "semiconductors," whose resistivities at room temperature fall between those of metallic conductors and those of insulators. The order of magnitude of these resistivities is as follows:

Resistivity in Ohm-cm

Metals	Semiconductors	Insulators
10^{-5}	10^{-2}–10^{6}	10^{6}–10^{15}

These semiconductors are usually metallic oxides, or other types of metallic compounds, which may be sintered in some inert binder such as clay. They are charac-

terized by their extreme sensitivity to voltage or temperature changes. They may be divided into three classes.

63. Low-voltage Varistors—Rectifiers. The classic examples of these are the selenium, copper-oxide, germanium, and silicon rectifiers and transistors. These are used primarily at low voltages and are characterized by their asymmetrical voltage-resistance curve; *i.e.*, the resistance is dependent on the direction of the applied voltage. They have extremely high resistance to negative voltages and low resistance to positive voltages (see Fig. 37*a*). This property is responsible for their rectifying ability. It should be noted that this is a characteristic of single-contact varistors, *i.e.*, varistors consisting of a single crystal or disk making one contact to the other electrode.

Some of the more typical applications for these voltage-varying resistors are for voltage regulation, signal clipping, balanced modulators, contact protection, and meter-overload protection.

64. High-voltage Varistors. A typical example of this class is the silicon-carbide resistor (thyrite). This, in contrast to the single-contact varistor described above, consists of a large number of silicon carbide granules sintered in a vitreous ceramic. Unlike the rectifier type, the resistance-voltage characteristic is symmetrical, *i.e.*, is independent of polarity (see Fig. 37*b*). This is explained partly by the fact that, although any single contact in the resistor mass may rectify, random distribution of large numbers of contacts in series and parallel results in equal numbers of contacts rectifying in opposite directions.

The characteristics of the high-voltage resistor depend upon the composition of material and heat-treatment. Changes in resistance of the order of 100,000 to 1 may be obtained by changes in the heat-treatment. Resistance changes of the order of 100,000 to 1 or more can be obtained by voltage changes of the order of 1,000 volts. Figure 38 shows a variety of characteristics possible with this type of resistor.

FIG. 37. Resistance-voltage characteristic of varistors: (*a*) low-voltage copper-oxide; (*b*) high-voltage silicon-carbide varistor.

The temperature coefficient is negative and a function of voltage, being higher at low voltages. At low voltages a change of 45°C may change the resistance value by a factor of 2.

65. Thermistors. This class of resistor consists of mixtures of one or more metallic oxides, such as ferric oxide, Fe_2O_3, nickel oxide, NiO, manganese dioxide, MnO_2, sintered in inert binders, such as clay. These resistors are temperature-sensitive and are characterized by an extremely high negative T.C. The explanation in simplified form is that at low temperatures semiconductors have relatively few "free" or conduction electrons, most of them being bound to their respective atoms; hence the resistivity is high. As the temperature is increased more and more of the electrons in bound states acquire energies sufficient to carry them over into "free" energy states, where they can be accelerated by an electric field; hence the resistance decreases, which accounts for the high negative T.C.

There are several methods of controlling the resistance, the more prominent of which are as follows:

1. *External heating*, or controlling the ambient temperature
2. *Direct heating*, in which the circuit current passes directly through the thermistor, thus heating the thermistor and altering its resistance

3. *Indirect heating*, in which the thermistor is placed inside or near a heating coil that carries the control current, and the heat thus generated controls the thermistor resistance

66. Static Self-heat Thermistor Characteristics (Voltage-Current). The current-voltage relationship in thermistors is unique. Ohm's law is not always obeyed since thermistor resistance will vary when the current reaches a point where self-heating begins. Figure 39 illustrates this. For small currents, the power developed is too small to cause appreciable self-heating and the resistance remains constant. The voltage-current relationship thus follows Ohm's law. As the current is increased, however, the power developed becomes sufficient to heat the thermistor and resistance

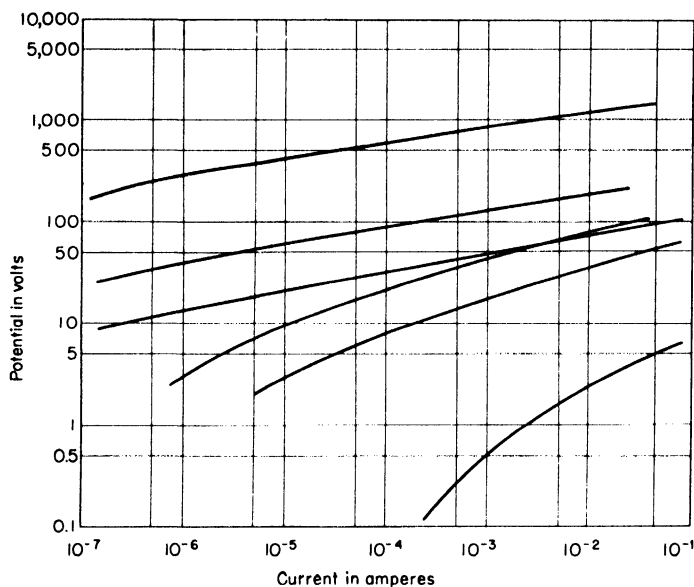



FIG. 38. Characteristics obtainable in silicon-carbide varistors.

decreases. The voltage is less than it would have been had the resistance remained constant. At some current the voltage will reach a maximum value and will then remain constant or decrease as current is increased. This relationship is considered in a static condition, allowing the voltage to reach an equilibrium value after each current increase.

A typical static characteristic is shown in Fig. 39. At some value of current the voltage reaches a peak value, known as the self-heat voltage; beyond this point the curve has a negative slope. Although the specific curve shapes vary for each type of material, they all show the same general form. An interesting feature of this curve is the falling voltage characteristic with increasing current to the right of the maximum voltage point. This shows a negative resistance characteristic, similar to that displayed by an arc, such as the Poulsen arc, indicating possible uses as an oscillator.

67. Dynamic Self-heat Thermistor Characteristics (Current-Time). The resistance variations resulting from self-heating are not instantaneous. Because of thermal inertia a certain interval of time elapses before the equilibrium current of a thermistor circuit is attained. When a voltage is applied across a thermistor there is a delay in reaching the equipment current. The initial current is determined by the unheated resistance of the element. As the thermistor heats up because of I^2R power, its resistance decreases and current increases. A build-up of current takes place until it is limited by the circuit resistance. Where necessary, a load resistance is placed in

series with the thermistor to limit the current to such a value that the I^2R power will not heat the element above its maximum recommended temperature. No load resistance is used where the applied voltage does not heat the thermistor above the recommended temperature.

A typical dynamic characteristic is shown in Fig. 40.

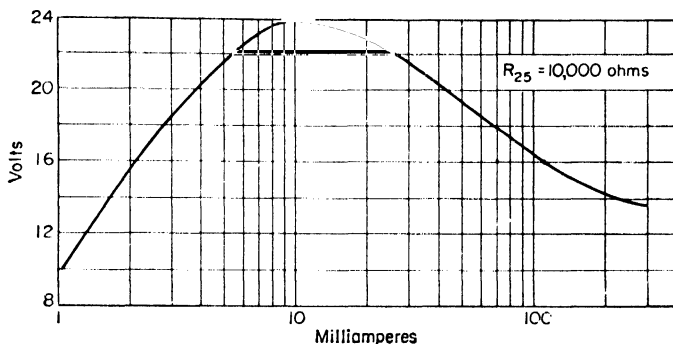



FIG. 39. Voltage-current characteristic of iron-oxide thermistor.

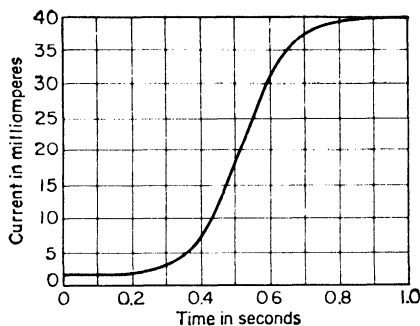



FIG. 40. Time characteristic of directly heated thermistor.

68. Applications of Negative-temperature Coefficient Resistors. By virtue of their high negative-temperature coefficients of resistance, their time constants, and their peculiar volt-ampere characteristics, thermistors are useful in a number of ways, a few of which are mentioned here:

1. *Resistance Thermometers.* A thermistor in series with an ammeter and constant-voltage source will indicate temperature when the ammeter is properly calibrated. Simple refinements of this system make it possible readily to indicate changes in temperature as small as 0.0005°C.

2. *Temperature Controllers.* Related to their use as thermometers, thermistors can be employed as temperature controllers.

3. *Time-delay Devices.* Owing to the thermal inertia of thermistors, the current through them will increase for a while after voltage is applied, as the thermistor heats up and the resistance decreases. This fact can be made use of to provide time delays ranging from a few milliseconds to several minutes.

4. *Motor Starters and Overcurrent Protection.* The high initial cold resistance of a thermistor, which decreases rapidly as it is heated by current flowing through it, is useful in motor protection or in other devices that draw high initial starting or heating-up currents.

5. *Temperature Compensation.* By proper choice and connection of thermistors the positive temperature coefficient of resistance of many passive networks can be neutralized to give a resultant network whose resistance is substantially independent of temperature over wide limits.

6. *Wattmeters.* Thermistors have been used to measure power, especially in the u-h-f band, by measuring their change in d-c resistance as they absorb all, or a given fraction, of the power to be measured.

7. *Manometers.* Since the dissipation constant of a thermistor depends on the thermal conductivity of the surrounding medium and this in turn depends, in the case of gases, on the pressure, thermistors can be readily used as sensitive pressure gages for gases. Through a similar mechanism they can also be used as velocity meters for gases or liquids.

8. *Automatic Gain Control.* By incorporating a thermistor in the feedback circuit of a negative feedback amplifier, the amount of feedback voltage can readily be varied automatically to compensate for changes in the output signal. Many types of such AGC circuits have been devised.

9. *Amplifiers, Oscillators, and Modulators.* Since thermistors can be made to display negative resistance characteristics, they can, theoretically at least, be used as oscillators, amplifiers, and modulators. Such devices actually have been made in laboratories and have functioned well at low and audio frequencies, but there are still many difficulties that prohibit their practical use at this time.

RESISTOR STANDARDS AND SPECIFICATIONS

There is considerable activity in the radio industry toward establishing standards and specifications for the performance of all components including resistors. Prior to the Second World War, RETMA had a program of standards under way which were aimed at satisfying commercial requirements. The advent of the war introduced conditions of operation encountered by the military so far outside the commercial orbit that new military specifications for components were set up. For example, ambient temperatures between −50 and 100°C and even higher had to be considered. Operation at altitudes in excess of 30,000 ft with the resultant low barometric pressure placed more severe requirements on components. As a result, two sets of standards, commercial and military, exist for components. However, developments in the components field are in the direction to make these two sets of standards converge.

69. Resistor Tests. Performance specifications and tests have become so extensive that it is not possible to give them in detail. The following is a list of standard tests, commercial and military, to which resistors are subjected. The detailed procedures can be obtained from the specific standard specifications listed.

Standard Tests Applied to Resistors

Test	Applied to
Resistance measurement	All types of resistors
Power rating by measurement of temperature rise	Fixed power resistors
Voltage coefficient	Fixed composition resistors
Temperature-resistance characteristics between −55 and 105°C	All resistors except variable composition
Rapid temperature cycling between −55 and +85°C	All resistors except fixed power wire-wound resistors
Dielectric strength or voltage breakdown	All resistors
Vibration	All resistors
Noise (rotational)	Variable composition resistors and precision potentiometers
Life test at rated load at specified ambient	All resistors except fixed power wire-wound
Short-time overload	All resistors
Humidity: 90 to 95% RH at 40°C, or humidity-load cycle at 25 to 65°C	All resistors
Salt-water cycling	External-meter resistors, precision wire-wound resistors and low-power wire-wound resistors
Thermal shock	Fixed power resistors
Wear test	All variable resistors
Low-temperature exposure	Fixed film (high-stability) composition resistors, fixed and variable
Terminal strength	All wire and lug types except on variable resistors

70. Resistor Specifications. The industry or commercial specifications are prepared by the engineering division of The Electronic Industries Association (EIA) and can be obtained from the EIA Engineering Office.[1] The military specifications

[1] 11 West 42d St., New York 36, N.Y.

for resistors are the Joint-Army-Navy-Airforce Specifications (abbreviated MIL) prepared by (and also obtainable from) the Armed Services Electro-Standards Agency (ASESA).[1] The list below gives the specifications now in existence.

Resistor type	Military specification	Commercial RETMA
Fixed, composition (insulated)	MIL-R-11B	RS-172
Fixed, wire-wound (accurate)	MIL-R-93A	
Fixed, film (high-stability)	MIL-R-10509B	
Fixed (composition film, very high frequency)	MIL-R-10683A	
Fixed, wire-wound (low-power)	JAN-R-184	REC-117
Fixed, non-wire-wound (power type)	MIL-R-11804B	
Fixed, wire-wound (power type)	MIL-R-26C	RS-155
External meter (high voltage, ferrule terminal)	JAN-R-29	
Variable, composition	JAN-R-94A	REC-121B
Variable, wire-wound (low operating temperature)	JAN-R-19	
Variable, wire-wound, precision	Proposed MIL-R-12934A	
Variable (wire-wound, power type)	MIL-R-22A	TR-133

References

Becker, Green, and Pearson: Properties and Uses of Thermistors, *Electrical Engineering*, November, 1946.
Driver-Harris Company: *Bulletin* R-40.
EIA Standards and Engineering Information.
Field, R. F.: "Boella Effect," unpublished paper.
Henney and Walsh: "Electronic Components Handbook," McGraw-Hill, 1957.
Howe, G. W. O.: Editorial, *Wireless Engineer*, April, 1934.
Nat. Bur. Standards Circ. 74.
"Potentiometer Handbook," Technology Instrument Corporation, Acton, Mass., 1954.
"Precision Resistors and Their Measurement," *Nat. Bur. Standards Circ.* 470.
"Resistance Handbook," Wilbur B. Driver Company.
Resistor Standards, MIL-R and JAN-R Standards.
Stein, S. J., and J. Riseman, Evaporated Metal Film Resistors, *Proceedings* 1954 *Electronic Components Symposium*.
"Thermistor," *Bulletin*, Western Electric Company.
"Varistors," articles reprinted from *Bell Laboratory Record*.

Catalogues:
Corning Glass Works.
Wilbur B. Driver Company.
Driver Harris Company.
General Electric Company, Carboloy Division, Thermistor Catalogue.
International Resistance Company.

[1] Fort Monmouth, N.J.

CHAPTER 3

INDUCTANCE;[1] MAGNETIC MATERIALS[2]

By F. G. Webber and Raymond L. Sanford

Part 1: INDUCTANCE

1. Definitions and Units. Inductance is that property of an electrical circuit whereby changes in the current flowing through it produce changes in the magnetic field associated with the current such that a counter emf is set up in that same circuit or in neighboring ones. If the emf is set up in the same circuit, the phenomenon is called "self-inductance"; if in a neighboring circuit, it is called "mutual inductance." Inductance is an effect exactly as inertia is an effect in a mechanical system. Inertia is that fundamental property of matter which causes a body to resist a change in its state of rest or motion. The important points are (1) that a change is necessary and (2) that the effect is to resist the change.

Inductance is measured by the emf produced in a conductor by unit rate of variation of the current. The practical unit of inductance is the *henry* (h). The henry is that inductance in which an induced emf of 1 volt is produced when the inducing current is changed at the rate of 1 amp per sec.

Since the henry is a rather large unit, especially in radio engineering work, it is commonly subdivided into two smaller units: a one-thousandth part, the millihenry (mh) = 10^{-3} henry; and a one-millionth part, the microhenry (μh), = 10^{-6} henry.

2. The Inductor. A straight conductor carrying a current has the property of inductance and, as will later be shown, is sometimes used as an inductor. However, in other configurations this property can be enhanced to become more useful. The most common inductors are coiled conductors. Since inductance is a magnetic effect, coiling the conductors gives a more concentrated field pattern. Flux set up by one turn of the coil links mutually with other turns of the coil, thereby increasing the total inductive effect. It follows that the closer the turns, the greater is the inductance. The design of an inductor would seem to be a relatively simple matter. However, it is not possible to construct an inductor capable of pure and simple inductance. (1) There is no perfect conductor and, therefore, losses in the form of resistance are introduced. (2) Distributed capacitance exists between turns of a coil and between the conductor and surrounding objects. The actual inductor is always a combination of resistance, inductance, and capacitance. The true inductance is the absolute or theoretical inductance. The apparent inductance is the effective inductance at any frequency, *i.e.*, the inductive minus the capacitive effect.

3. Resistance Effects. The reason for the difference between the d-c and the r-f resistance of a conductor is the so-called "skin effect." Consider two small filaments of a conductor, one filament at the center and one near the surface of the conductor. All the flux produced by the current through the conductor will link the center filament, but the flux within the conductor will not link the filament near the surface. This means that the center filament will have a higher self-inductance than the surface filament. This higher self-inductance means a higher reactance path for the current in the conductor. More current, therefore, flows at the conductor surface. The redistribution of the current forces more current through less of the conductor, giving higher losses and more heating. The r-f resistance can become several times the d-c resistance. See Sec. 8, Chap. 2, for ratios of a-c to d-c resistance.

[1] By F. G. Webber, F. W. Sickles Div., General Instrument Corp.
[2] By Raymond L. Sanford, National Bureau of Standards.

Since the smaller the conductor, the less the skin effect, if the conductor is broken down into many small strands, twisted about ten turns per foot and each strand insulated from the others, the skin effect is lessened. Such wire is known as "litz" wire. Skin effect is a function of the following factors:

$$d\sqrt{\frac{\mu f}{\rho}}$$

where d = diameter of conductor
μ = permeability of conductor
ρ = specific resistance of conductor
f = frequency

For air-core inductors, skin effect is the major consideration; for iron-cored or copper-cored inductors eddy-current and hysteresis losses in the core reflect as resistances in the coil. Both these losses vary with frequency.

The eddy-current loss for any material may be expressed by

$$W_e = eVt^2f^2B_{max}^2$$

where e = loss per unit volume
V = volume
t = thickness of laminations or particle size
B_{max} = maximum flux density
f = frequency

The hysteresis loss for iron, according to the Steinmetz empirical law, is

$$W_h = nfVB_{max}^{1.6}$$

where n = loss per unit volume
V = volume
B_{max} = maximum flux density
f = frequency

The total loss expressed in watts in an iron-cored inductor is

$$W = I^2R_{ac} + W_h + W_e$$

Another factor affecting the resistance of an inductor comes into play when that inductor "couples" with other circuit elements. If current is induced in the coupled circuit, losses occur and the resistance is increased. The general definition of effective, equivalent, or apparent resistance is such that this resistance multiplied by the square of the effective value of the current gives the true average rate at which heat energy is dissipated.

4. Distributed Capacitance Effects. The distributed capacitance can be assumed to be a capacitance shunted across the inductance. The effect is to give an apparent inductance that is less than the true inductance. The higher the distributed capacitance, the greater is the ratio of true to apparent inductance. Furthermore, because this distributed capacitance has dielectrics other than air in many cases, other losses can be introduced as dielectric losses. See the chapter on Capacitance for a further explanation of such losses.

The distributed capacitance of a multilayer coil, *e.g.*, a universal winding, is a highly complex thing. Capacitances occur between layers as well as between turns. The result is a network of capacitors in series, in parallel, and in series-parallel, which may cause the total distributed capacitance to increase or decrease as the inductance increases. This effect is shown in Table 2, which gives the approximate distributed capacitance of choke coils. This also explains one reason why the Q of a coil may increase with inductance up to a point and then begin to decrease.

The distributed capacitance of a single-layer solenoid is a simpler function and is approximately proportional to the radius of the form and independent of the number of turns and the winding length for solenoids of the same inductance.

5. Figure of Merit, or Q. To judge the quality of an inductor, which of necessity cannot be perfect, a figure of merit is used. This figure of merit, or Q, gives an idea of the ability of an inductor to perform its primary function, that of supplying inductance at some given frequency. Q is defined as the ratio of inductive reactance to effective resistance. Therefore, the higher the true inductance, the lower the distributed capacitance, and the lower the losses from all sources, the higher the Q. Q varies with frequency since both the reactance and the losses vary with frequency.

FIG. 1. Effect of Q on sharpness of resonance.

Power factor is the ratio of impedance to resistance for any circuit. Since, for all practical inductors, the impedance and the reactance are nearly the same, the power factor is very nearly the reciprocal of Q.

$$Q = \frac{X}{R} = \tan \phi$$

$$\text{Power factor} = \frac{R}{Z} = \cos \phi$$

where ϕ is the angle between Z and R in a right-angled triangle with a hypotenuse of Z. For very large angles, the cosine is approximately equal to the reciprocal of the tangent, the cotangent. If the Q of the inductor is 10 or more, its power factor will be within ½ per cent of the reciprocal of the Q.

FIG. 2. Q vs. frequency for various types of windings: curve 1, two-layer bank winding; curve 2, progressive universal winding; curve 3, solenoid winding; curve 4, universal winding. All coils are 250 μh, wound with $^{7}/_{41}$ SSE wire on $^{7}/_{8}$-in.-OD phenolic forms.

Q has another significance, that of showing the relative sharpness of resonance of series-resonant circuits. If the absolute value of the admittance of the circuit is plotted against frequency, a peaked resonance curve is obtained. If the resistance is increased, the admittance will be decreased at all frequencies but much more markedly near the resonant frequency. The curve will be much flatter under these conditions. The less the resistance, the steeper and sharper the curve will be, resulting in a more selective circuit. Since the inductor has the greatest part of the resistance of the circuit, the Q of the inductor is a good indicator of the expected selectivity of the circuit. High Q means good selectivity.

The curves in Fig. 1 show the sharpness of resonance for both good and poor Q windings. The curves in Fig. 2 show the Q vs. frequency characteristics for several types of windings.

MEASUREMENTS OF INDUCTANCE

The story of the various methods of measurements to be discussed here is more thoroughly covered in the chapters on electrical measurements. However, the practical problems involved in measurements concerned with inductors will be considered.

6. Apparent and True Inductance. Apparent inductance may be measured by finding the capacitance necessary to resonate the inductor to a given frequency. Such

measurements are frequently made since standard calibrated capacitors and frequency sources are usually handy in the laboratory. The apparent inductance is then calculated from the formula

$$L = \frac{1}{4\pi^2 f^2 C}$$

where L is in henrys, f in cycles, C in farads, or

$$L = \frac{2.533 \times 10^4}{f^2 C}$$

where L is in microhenrys, f in megacycles, C in micromicrofarads.

Apparent inductance matching is used in the production testing of inductors when the beat-frequency or single-frequency method of testing is employed. In this case the inductor under test is compared with a standard or set-tested inductor by comparing the amount of capacitance required to tune each inductor to the same frequency.

True inductance measurements are approximated on the 1,000-cycle impedance bridge since, for most inductors used at radio frequencies, the reactance of the distributed capacitance is very high at 1,000 cycles and in parallel with the comparatively low inductive reactance it has little effect. The effective resistance is taken care of in the bridge balance.

True inductance may also be measured by finding the amounts of capacitance required to resonate the inductor to two different frequencies and then calculating the inductance by the following formula:

$$L = \frac{n^2 - 1}{4\pi^2 f_1^2 (C_2 - C_1)}$$

$$n = f_1/f_2$$

where f_1 and f_2 = the two frequencies

C_1 and C_2 = capacitances required to resonate at f_1 and f_2

The effect of distributed capacitance is eliminated as is shown by the derivation of the general true inductance formula:

$$f_1 = \frac{1}{2\pi \sqrt{L(C_1 + C_d)}} \tag{1}$$

$$f_2 = \frac{1}{2\pi \sqrt{L(C_2 + C_d)}} \tag{2}$$

$$n = \frac{f_1}{f_2} \quad \text{or} \quad f_1 = nf_2 \tag{3}$$

Substituting Eq. (3) in Eq. (1),

$$n^2 f_2^2 4\pi^2 = \frac{1}{L(C_1 + C_d)} \tag{4}$$

Substituting Eq. (2) in Eq. (4),

$$\frac{n^2}{L(C_2 + C_d)} = \frac{1}{L(C_1 + C_d)}$$

whence

$$C_d = \frac{C_2 - n^2 C_1}{n^2 - 1} \tag{5}$$

Substituting Eq. (5) in Eq. (1),

$$f_1^2 4\pi^2 = \frac{1}{L\left(C_1 + \frac{C_2 - n^2 C_1}{n^2 - 1}\right)}$$

whence

$$L = \frac{n^2 - 1}{4\pi^2 f_1^2 (C_2 - C_1)}$$

which is the general true inductance formula.

Example: If $f_2 = 600$ kc, $f_1 = 1{,}200$ kc, $C_2 = 130$ μμf, $C_1 = 30$ μμf, then

$$n = \frac{1{,}200}{600} = 2$$

$$\begin{aligned} L &= \frac{2^2 - 1}{4\pi^2(1.2 \times 10^6)^2(130 - 30)10^{-12}} \\ &= 5.28 \times 10^{-4} \text{ h} \\ &= 528 \ \mu\text{h} \end{aligned}$$

and

$$\begin{aligned} C_d &= \frac{C_2 - 4C_1}{3} \\ &= \frac{130 - (4 \times 30)}{3} = 3.33 \ \mu\mu\text{f} \end{aligned}$$

For simplicity of calculation and equipment design, it is common to choose the frequencies as the fundamental and the second harmonic so that

$$\frac{f_1}{f_2} = 2 \qquad L = \frac{3}{4\pi^2 f_1^2 (C_2 - C_1)} \qquad \text{or} \qquad \frac{0.07599}{f_1^2(C_2 - C_1)}$$

Note also that this gives a method for the measurement of distributed capacitance from the general formula (5) $C_d = \frac{C_2 - n^2 C_1}{n^2 - 1}$ or, for the more usual case where $n = 2$, $C_d = \frac{C_2 - 4C_1}{3}$.

True inductance measurements by this method are sometimes used but are generally spurned in favor of the simpler 1,000-cycle bridge measurements.

True inductance or two-frequency inductor matching is often used in the production testing of inductors to ensure more accurate "tracking" of the r-f part of single-dial radio sets.

7. Two-frequency Test. The test equipment (Fig. 3) consists of two oscillators, one used as a reference and the other used with the inductor under test. The signals

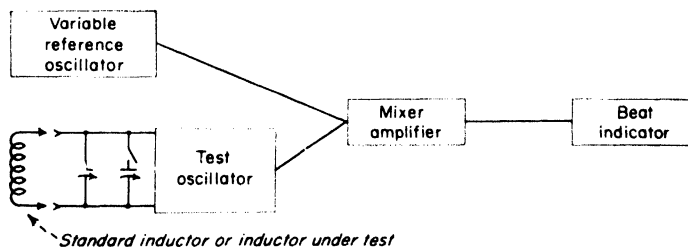



FIG. 3. Two-frequency or true-inductance matching setup.

from the two oscillators are mixed, fed into an amplifier and thence to either an audible or a visual beat indicator to show when the two oscillators are at the same frequency. The reference oscillator is set to the lower of the two frequencies (in the example, 600 kc). A standard inductor, either tested in a radio set or accurately measured as to its true inductance, is placed in a test jig so that it forms the inductance for the

second oscillator circuit. A variable capacitor is shunted across the inductor. This capacitor is sufficient to tune the inductor to the second harmonic of the reference oscillator (in the example, 1,200 kc). A switch is provided that allows approximately three times that capacitance to be added in parallel when desired. With the additional capacitance, the inductor will tune to the fundamental frequency (600 kc). The standard inductor is tuned to the second harmonic and then, without disturbing this capacitor, the second capacitor is added and with it the inductor is tuned to the fundamental frequency. Both capacitor readings are noted. An unknown inductor may now be matched to the standard by removing the standard from the test jig and inserting the other inductor. That inductor is then adjusted until both capacitor readings are within the desired tolerance of the standard. The difference in capacitance readings between the unknown and the standard at the second harmonic gives an indication of the differences in their distributed capacitances, and the difference in readings at the fundamental frequency gives the difference in true inductance between the two.

8. Measurement of Distributed Capacitance. One very common method of measuring distributed capacitance has already been discussed under measurement of inductance. That method must be used with caution since it depends upon the difference of two numbers of much greater magnitude than the resulting distributed capacitance and is, therefore, liable to introduce large errors.

It is possible to determine the natural resonant frequency of an inductor by placing it in parallel with another resonant circuit whose resonant frequency is calibrated and can be varied. The frequency of the calibrated circuit is varied until the introduction of the inductor being measured does not affect the tuning but only the circuit Q. This is the natural resonant frequency of the inductor since at that point it has no reactance but acts as a pure resistance loss. Assuming that the inductance remains relatively constant over a range of frequencies, the distributed capacitance can then be found by measuring the capacitance required to tune the inductor to some lower frequency, far removed from the natural resonant frequency, and then calculating the distributed capacitance by the formula

$$C_d = \frac{f^2 C}{f_n^2}$$

where f = lower frequency
C = capacitance to tune to that frequency
f_n = natural resonant frequency

In practice f_n is about ten times f. This method is more accurate than the previous method, especially for distributed capacitances of less than 10 $\mu\mu$f.

9. Measurement of Q. The figure of merit, Q, is most readily measured on Q meters that have been calibrated to read directly over large frequency ranges. The Q meter is based on the fact that the ratio of the voltage developed across the capacitor in a resonant circuit to the voltage introduced into the circuit is equal to Q. Thus a measured current from an oscillator flows through a resistor of very low resistance in series with the inductor to be measured, introducing known voltage into the circuit. A built-in calibrated capacitor completes the series circuit. A voltmeter, calibrated directly in Q, measures the voltage developed across the capacitor. This voltmeter has negligible power consumption.

In resonant circuits containing a well-designed, low-loss capacitor, the resistance of the capacitor is negligible as compared with the resistance of the inductor. The total losses can be considered as occurring in the inductor alone. Under these conditions,

$$Q = \frac{f_r}{f_2 - f_1}$$

where f_r is the resonant frequency, f_1 and f_2 are the frequencies above and below resonance at which the power dissipated is one-half the power dissipated at resonance, and

the current is 0.707 times the current at resonance. Also

$$Q = \frac{C_r}{C_2 - C_1}$$

where C_r is the capacitance at resonance, and C_1 and C_2 are the capacitances corresponding to frequencies f_1 and f_2. With these formulas, Q measurements may be made with equipment usually available around a laboratory.

10. Measurement of Mutual Inductance. The simplest method of measuring mutual inductance is the Felici balance in which the unknown mutual inductance is compared with a variometer whose mutual inductance with respect to rotation has been calibrated. The primaries of both inductors are connected in series across an a-c source, usually 1,000 cycles. The secondaries are also connected in series, so that the voltages induced in them oppose each other, and in series with a null detector, usually a pair of headphones with a suitable amplifier. At the null, the mutual inductance of the unknown is equal to the mutual inductance of the known calibrated variometer.

If an inductance bridge is available, the mutual inductance, M, may be found by measuring the total inductance of the primary and secondary inductances in series aiding and then in series bucking. M is calculated from the formula

$$L_M = \frac{L_{\text{aiding}} - L_{\text{bucking}}}{4}$$

Coefficient of coupling is a term sometimes encountered in working with coupled inductors. This coefficient, K, is the ratio of the mutual inductance to the geometric mean of the primary and secondary inductances, or

$$K = \frac{L_M}{\sqrt{L_p L_s}}$$

TYPES OF INDUCTORS

11. Straight Conductors. Since any conductor carrying a current has the property of self-inductance, the simplest form of inductor is a straight wire. Straight wires are used as inductors in h-f applications where the inductance must be very low. In some cases the inductance is furnished by the current flowing along the plates of the variable tuning capacitor. At the frequencies at which straight conductors become important as inductors, many other complications arise not within the scope of a discussion of inductors; therefore, they are simply mentioned here without any detail.

For reference, a straight piece of No. 10 AWG wire 1 in. long has an inductance of about 0.015 μh at low frequencies, if the return conductor is assumed to be remote.

FIG. 4. Solenoid with spaced section for adjusting the inductance.

12. Single-layer Solenoid Windings. The single-layer solenoid winding is probably the most important type of inductor for frequencies from about 1 to 100 Mc, or higher. It has the advantage of being fairly simple. Its inductance can be varied by spacing a fraction of the total turns away from the main body of the winding (Fig. 4). Then by moving a few of the spaced turns nearer to the main body, the inductance can be raised slightly. By moving a few turns away from the main body to the spaced portion, the inductance can be lowered slightly. Such adjustment of inductance is important in the tracking of the r-f portions of superheterodyne receivers.

The single-layer solenoid winding has the further advantage of low distributed capacitance, since the first and last turns which have the greatest voltage developed between them are relatively far apart. The voltage between any two turns is only $1/n$ of the total voltage if n is the total number of turns.

This type of winding can be made to have a fairly constant inductance with temperature changes since it can be wound in grooves in the coil form or firmly cemented to the form.

For a given design, Q variations of ± 5 per cent and inductance variations of ± 2 per cent can be expected in the normal production of solenoids. In the case of coils used in broadcast receivers whether the windings are solenoids, universals, progressive universals, or banks, the common commercial inductance tolerances after adjustment are $\pm \frac{1}{2}$ per cent for oscillator coils and ± 1 per cent for r-f and antenna coils. See Table 1 for a list of common h-f solenoids.

Table 1. High-frequency Solenoids

OD, in.	Wall, in.	Wire	Tpi*	Turns	Inductance, μh	Q	Frequency, Mc
1	3/64	20E	20	5 9/16	1.1	105	7
1	3/64	20E	20	5 1/4	0.9	95	8
7/8	1/32	22E	22	7 1/2	1.6	165	12
3/4	1/32	22T	22	10 1/2	1.7	146	7
3/4	3/64	22T	22	7 1/2	1.1	165	14
3/4	1/16	20E	19	11 1/8	2.1	127	7
3/4	1/16	20E	20	9 1/2	1.2	165	14
3/4	1/16	22E	16	12 1/2	2.2	115	10
3/4	1/16	22T	22	7 1/2	1.1	165	14
9/16	1/32	24E	30	15	2.7	145	10
9/16	1/32	24E	30	12 1/4	1.9	137	10
1/2	1/16	22E	34	19 1/8	3.0	133	12
1/2	1/16	22E	36	24 1/2	4.3	112	7

* Tpi is turns per inch for grooving.
NOTE: All forms are phenolic, grooved.

13. Multilayer Solenoids. Multilayer solenoids are usually wound with the first layer running from left to right, the next running from right to left, etc. (Fig. 5). If the layers are separated by paper, the winding is known as a "paper-layer" solenoid. This type of winding has the disadvantage of high distributed capacitance since the beginning of one layer and the end of the next layer, which are at high potential difference, are directly above one another. Adding paper between the layers serves to reduce the distributed capacitance, improve the voltage breakdown between layers, and provide a smooth winding surface for the next layer. Multilayer solenoids are used at power and audio frequencies but never at radio frequencies where distributed capacitance is important.

FIG. 5. Cross section of multilayer solenoid winding.

14. Universal Windings. The universal winding is a comparatively narrow multilayer winding in which the conductor spirals back and forth across the width of the winding as the winding is built up. This results in a self-supporting lattice type of construction. By changing the ratio of the pitch of this spiral to the number of turns, varying numbers of crossovers per turn and turns per layer can be obtained. The distributed capacitance is fairly low since turns with high potential difference simply cross each other and do not lie parallel to one another for appreciable distances. Large inductances can be efficiently put into small spaces through the use of universal windings. They are commonly found in r-f chokes, i-f transformers, high-impedance primaries, etc.

By making the universal winding in several sections of pi's side by side on the form,

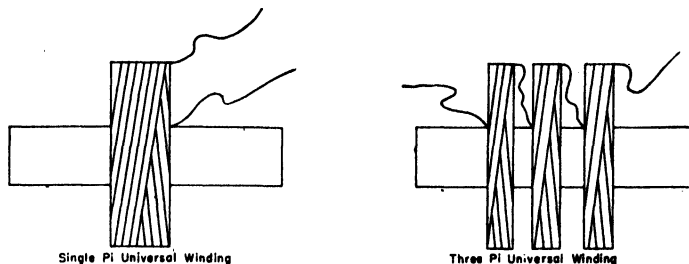



FIG. 6. Single- and three-pi windings.

the distributed capacitance can be further reduced and, since the field pattern is elongated, the effect of placing a shield around the winding is diminished. The effective diameter is also decreased, a fact that can be important when powdered-iron cores are used. Honeycomb coils were a type of universal winding with relatively few, widely spaced turns per layer giving a typical "honeycomb" appearance. The universal winding can be varied in inductance somewhat for tracking purposes. It can be reduced by squeezing, which shortens the effective diameter. The inductance may

Table 2. Typical Universal Wound Choke Coils

All coils are two-cross, single-pi universals. Coils of 0.5 to 20 mh are ¼ in. wide wound with No. 36 SSE wire. Coils of 25 to 70 mh are wound with No. 38 SSE wire and are 5/16 in. wide. All values are approximate.

Inductance, mh	Turns	OD, in.	D-c resistance, ohms	Distributed capacitance, μμf
0.5	185	0.60	12	11.3
1.0	260	0.63	17	11.0
1.5	320	0.67	21	10.7
2.0	370	0.69	26	10.2
2.5	400	0.71	29	9.6
3.0	450	0.73	31	9.5
4.0	510	0.76	37	7.6
5.0	570	0.79	42	7.5
10.0	790	0.90	61	7.1
15.0	940	0.98	78	8.2
20.0	1,060	1.04	93	9.8
25.0	1,390	0.83	159	8.8
30.0	1,450	0.85	183	9.0
40.0	1,680	0.90	219	9.5
50.0	1,850	0.97	269	7.4
60.0	2,000	1.01	274	6.9
70.0	2,150	1.07	294	6.6

Table 3. Effect of Form Diameter on *Q**

Form diam, in.	*Q* in air	*Q* in 1¾-in.-square Al shield	*Q* in 1¾-in.-diam, round Al shield
⅜	125	118	114
½	132	122	117
⅝	131	120	116

* 1-mh coil, 7/41 SSE wire, ¼ in. wide, 2½ crosses per turn, single pi.

be raised by squeezing the thickness of the winding to increase its effective diameter. Such distortion of the winding is not too desirable since it makes for poor stability, reduces Q, and increases distributed capacitance. Pi windings may be moved closer together or farther apart to accomplish the same end with the same disadvantages.

Table 4. Effect of Width of Winding on Q*

Width, in.	Q in air
3/16	101
7/32	105
1/4	109

* 1-mh coil, 7/41 SSE wire, 2 crosses per turn, single pi.

Universal windings are commonly used for inductors from 10 μh to 1 h. They may have Q's up to about 150. Common Q variations found in commercial windings of the same design and inductance are ±10 per cent. The usual inductance tolerance for unadjusted windings is ±5 per cent. See Tables 2 to 7 for data on universal windings.

An article by A. W. Simon[1] covers the general theory of universal windings and gives other references to work done on the mathematical study of these windings.

15. Bank Windings. Banked or bank windings are another form of multilayer windings in which a special arrangement of the turns in the various layers tends to reduce distributed capacitance. The

Table 5. Effect of Number of Crosses per Turn on Q*

No. crosses	Q in air	Q in 1¾-in. square Al shield	Q in 1¾-in.-diam round Al shield
1¼	120	113	110
2	109	102	99
2⅕	125	118	114

* 1-mh coil, 7/41 SSE wire, ¼ in. wide, ⅜-in.-diam form, single pi.

Table 6. Effect of Number of Pi's on Q*

No. pi's	Q in air	Q in 1¾-in.-square Al shield	Q in 1¾-in.-diam round Al shield
2	131	124	120
3	134	128	123
4	132	126	123

* 1-mh coil, 7/41 SSE wire, pi's 3/32 in. wide spaced 1/16 in., ½-in.-diam form, 4 crosses per turn.

Table 7. Effect of Wire Size on Q;* Relative Cost and Relative OD

Wire size	Q in air	Q in 1¾-in.-square Al shield	Q in 1¾-in.-diam round Al shield	Relative cost	Relative OD
3/40	103	97	95	1 (lowest)	1 (smallest)
5/40	119	108	105	5	5
5/41	121	113	110	3	2
6/42	130	120	116	2	1 (smallest)
7/41	133	120	114	4	3
10/41	139	122	107	7	7 (largest)
10/42	140	128	123	6	6
12/43	157	139	133	8	4
15/44	176	149	142	9 (highest)	5

* 1-mh coil, 3⅔ crosses per turn, 3/16 in. wide, ½-in.-diam form, single pi.

[1] *Proc. IRE*, January, 1945.

arrangement is shown in Fig. 7. This type of winding is difficult to produce and has, therefore, lost favor.

Commercial Q variation for a given design is ± 10 per cent while the inductance variation is about ± 5 per cent for unadjusted windings. See Table 9 for a comparison of bank and progressive universal windings.

Table 8. Effect of Broken Strands of Litz Wire on Q and D-c Resistance in a Universal Winding

No. of broken strands	Q	$R_{d\text{-}c}$, ohms
0	178	39
1	174	43
2	172	47
3	169	52
4	165	58
5	161	67
6	156	78
7	147	94
8	138	118
9	126	156
10	105	231
11	71	473

800 turns, ¼ in. wide, ½-in.-diam form, 12/43 SSE wire.

16. Progressive Universal Windings. To a large degree, progressive universal windings have supplanted the bank windings. This winding is produced in a manner similar to the universal winding except that instead of spiraling back and forth in the same area, the spiraling also progresses along the form and gives a winding that is spread out horizontally rather than built up vertically. A simple progressive winding looks somewhat like a bank winding. Because it is possible to vary the crossovers per

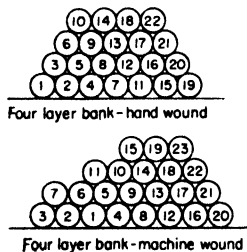



FIG. 7. Cross section of bank windings.

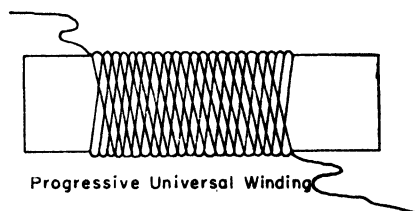



FIG. 8. Appearance of progressive universal winding.

turn as well as the rate of progression, a great variety of windings is possible. This type of winding helps make possible the high-Q permeability-tuned inductors to be discussed later.

The commercial variations in Q and inductance of a given design of a progressive universal winding are the same as those for universal and bank windings, ± 10 and ± 5 per cent, respectively.

An article by A. W. Simon[1] gives some of the theory of this complex winding type.

17. Other Types of Windings. There are a few other types of windings of more or less importance. For example, the flat spiral is widely used in loop antennas for broadcast receivers. Spider-web, basket-weave, and similar windings are cumbersome and difficult to produce. The toroidal winding has its conductor wound on the surface

[1] *Proc. IRE*, December, 1945.

Table 9. Comparison of Progressive Universal and Two-layer Bank Windings

Air inductance of all windings 250 μh; wire 7/41 SSE

Form diam, in.	600 kc		1,000 kc		1,400 kc	
	Prog.	Bank	Prog.	Bank	Prog.	Bank
			Q in air			
5/8	110	120	142	148	148	147
3/4	123	134	160	166	165	165
7/8	131	136	170	174	171	176
1	134	143	172	180	169	177
			Q in 1 13/16- by 1 7/8-in. Al shield			
5/8	82	106	110	135	119	141
3/4	80	106	121	137	130	143
7/8	79	101	120	130	130	137
1	84	99	113	128	122	135

of a toroid. When carefully made, the field is almost completely self-contained. It requires little shielding and is sometimes used in making inductance standards. When a high-permeability core is used, a maximum inductance can be obtained in a minimum space. Such toroids are finding increased use in the missile and computer fields.

18. Iron-cored Inductors. Since the permeability of the medium has a definite effect on the flux that causes self-inductance, it is obvious that the inductance of an air-cored winding can be increased by introducing material, such as iron, into the

Table 10. Toroidal Iron-core Coils

This table* gives data derived from measurements on toroidal cores of the most important iron powders available on the market. All toroids were prepared in the same way for reasons of comparison, *i.e.*, with 4 per cent bakelite (type 7095), 2 per cent Sterotex for lubrication, 50 tons per sq in. pressure, and 50 g weight. Their OD is 2.25 in., ID 1.50 in., and they are wound with a single layer of 200 turns of No. 25 magnet wire.

The data were obtained for the most part according to the method of V. E. Legg. *Bell System Tech. J.*, **15,** 39, 1936. Thus, the following units apply:

Density, g/cm³
Permeability, gausses gilbert/cm
Hysteresis loss coefficient, 10^{-3} ohm/h/gauss cycle
Eddy-current loss coefficient, 10^{-6} ohm/h/cycle squared
Magnetic stability, percentage change of permeability after 4 amp passage through winding

Material	Density	Permeability	Hysteresis	Eddy current coeff	Stability
Carbonyl L	5.7	24.8	3.1	0.13	+1.4
Carbonyl C	5.7	16.7	1.1	0.14	+0.6
Carbonyl E	4.9	10.4	0.3	0.11	+0.6
Carbonyl TH	4.9	9.6	0.3	0.10	+0.18
Carbonyl SF	4.8	8.1	0.3	0.10	+0.15
Electrolytic	5.6	23.4	2.4	0.33	−0.17
Hydrogen reduced, 1	5.6	18.4	2.6	0.12	+1.3
Hydrogen reduced, 2	5.5	16.9	1.0	0.12	+1.6
Hydrogen reduced, 3	5.2	12.5	3.1	0.11	+1.5
Magnetite, 1	4.1	7.9	9.1	11.5	−2.1
Magnetite, 2	4.0	5.7	6.8	0.21	−1.6
Magnetite, 3	3.5	3.1	0.3	0.085	+0.10

* All data from General Aniline and Film Corporation.

winding to increase the permeability of the medium. However, the introduction of the material also introduces a new source of losses. Magnetic losses due to hysteresis and eddy currents were discussed in a preceding section. At low frequencies, in the power or audio region, laminating the iron will substantially reduce the eddy-current losses. As long as large amounts of d.c. are not present in the inductor, which would tend to produce saturation of the iron and hence a nonlinear inductance as the current is varied, the only major problem is to get as much permeability as possible in order to get as much inductance as possible in a small space.

At radio frequencies it is not possible or practical to make laminations thin enough to reduce the losses to a sufficient degree. An entirely different technique is, therefore, employed in making the core: powder metallurgy. In this process the iron is in a very finely powdered state, produced either mechanically or chemically. The powder is combined with suitable binders, compressed, and sintered to produce a very compact mass which nonetheless consists of separate particles with an average size of 10 microns or less.

Carbonyl iron powders are obtained by thermal decomposition of iron pentacarbonyl. There are five different grades in production, designated as L, C, E, TH, and SF powder. Each of these is obtained by special process methods and has its special field of application. The particles making up the powders E, TH, and SF are spherical with a characteristic structure of increasingly larger shells. The particles of L and C are made up of homogeneous spheres and agglomerates.

The chemical analysis, the weight-average particle size as determined in a Roller air analyzer, the "tap density," *i.e.*, the density of the loose powder packed by tapping the container in a prescribed manner, and the apparent density or bulking factor as determined in a Scott volumeter are given in Table 11 for the five grades.

Table 11*

Grade	Chemical analysis, per cent			Wt—ave. diam, microns	Tap density, g/cm³	Apparent density, g/cm³
	Carbon	Oxygen	Nitrogen			
L	0.005-0.03	0.1 -0.2	0.005-0.05	20	3.5-4.0	1.8-3.0
C	0.03 -0.12	0.1 -0.3	0.01 -0.1	10	4.4-4.7	2.5-3.0
E	0.65 -0.80	0.45-0.60	0.6 -0.7	8	4.4-4.7	2.5-3.5
TH	0.5 -0.6	0.5 -0.7	0.5 -0.6	5	4.4-4.7	2.5-3.5
SF	0.5 -0.6	0.7 -0.8	0.5 -0.6	3	4.7-4.8	2.5-3.5

* All data from General Aniline and Film Corporation.

Table 12*

Grade	Effective permeability at 1 kc	Relative quality factor at				
		10 kc	150 kc	200 kc	1 Mc	100 Mc
L	4.16	100	96	90	43	1
C	3.65	94	100	98	72	3
E	3.09	81	94	100	97	30
TH	2.97	81	93	98	100	54
SF	2.17	62	71	78	84	100

* All data from General Aniline and Film Corporation.

NOTE: Q values actually measured can be obtained by multiplying the rows, respectively, by 0.78, 1.09, 1.25, 2.63, and 1.62.

L and C powders are also used as powder-metallurgical material because of their low sintering temperatures, high tensile strengths, and other very desirable qualities. Sintering begins below 500°C, and tensile strengths reach 150,000 psi.

With reference to the chemical analysis shown in Table 11, it should be noted that spectroscopic analysis shows the rest to be iron with other elements present in traces only.

Carbonyl iron powders are primarily useful as electromagnetic material over the entire communication frequency spectrum.

Table 12 gives relative Q values and effective permeabilities for the different grades of carbonyl iron powder. The values are derived from measurements on straight cylindrical cores placed in simple solenoidal coils. Although the data were not obtained at optimum conditions, the Q values as expressed in percentage of the best core give an indication of the useful frequency ranges for the different powder grades.

FIG. 9. Selectivity characteristics of two-stage 455-kc i-f amplifiers using 6A8, 6K7, 6K7, and 6Q7 tubes. Curve 1, air-trimmed, four-pi universal air-core units; curve 2, single-pi universal permeability-tuned units.

19. Variable Inductors. Probably the simplest form of variable inductor is a tapped winding in which the inductance is varied by taking more or less of the total available winding. This is a step-by-step method of variation and usually not satisfactory. Another form, which is rather bulky, difficult to manufacture, and now little used, is the variometer. This consists of two similar windings so disposed that they are closely coupled and can be changed from a mutual aiding to mutual bucking position in a continuous manner. This is accomplished by placing one winding on the inner surface of a sphere and the other on the outer surface of a slightly smaller sphere within it. The two windings are connected in series. Rotating one sphere with respect to the other gives the desired change in mutual and, hence, in total inductance.

Much more common is the variation of inductance through the variation of the permeability of the core. The effective permeability of the core for a particular inductor is simply the ratio of the inductance with the core in the winding to the inductance with an air core. The effective permeability may be greater than or less than unity, the air-core value. The commonly used windings are solenoid, universal, and progressive universal.

A slug of low-resistance conducting material such as copper will give an effective permeability as low as 1/1.4. This is the result of the eddy currents which are induced in the core. The eddy currents cause flux which always bucks the inducing flux and, hence, reduces the self-inductance.

Copper cores are used at frequencies above about 40 Mc. At such frequencies Qs are relatively high and the decrease in Q caused by the core is not serious. The use of copper cores allows a maximum conductor length since the inductance decreases as the core is inserted. This is mechanically desirable in h-f winding design.

A slug of powdered iron may have an effective permeability as high as 14. Pow-

dered-iron cores are usually used at frequencies lower than 100 Mc. At these frequencies high *Q*s are more important. With high permeabilities, minimum conductor length is required to produce the desired inductance. This means minimum copper losses and, consequently, high *Q*. However, high-permeability cores may have high inherent losses so that the highest over-all *Q* may not be realized with the highest permeability core.

By properly designing a tapered winding, the frequency to which it is tuned by introducing the iron core into the winding can be made linear with linear motion of the core. Furthermore, by the use of such variable-pitch windings, both solenoid and progressive universal types, and by use of movable iron cores, an r-f tuning unit of the two- or three-gang type may be made using permeability-tuned inductors rather than the usual gang capacitors. Such permeability-tuned units can be made small, light, and highly efficient. "Tracking" is accomplished by matched windings and iron cores.

In general, a circuit that contains a permeability-tuned inductor with a stable silvered-mica type of capacitor has far greater frequency stability with time and temperature than a circuit containing a fixed inductor and a variable mica capacitor.

Figure 9 shows the over-all selectivity characteristics of two two-stage 455-kc i-f systems. Curve 1 is that of an air-trimmed, four-pi universal-wound coil system. Curve 2 is that of a single-pi, universal-wound, permeability-tuned coil system. The tubes used in the i-f amplifiers were 6A8, 6K7, 6K7, 6Q7. Despite the fact that less expensive single-pi coils are used, the increased sharpness of resonance due to the use of iron-cored, permeability-tuned units is obvious.

20. Inductor Design. The design of inductors is, for the most part, somewhat more of an art than a science. Many practical factors influence it to such an extent that the theoretically best inductors are not obtainable. Available space, shielding requirements, materials of construction, and cost are usually the controlling factors. Where they are practical, the following formulas for the calculation of inductance are given, but for the most part some examples of the effects of the various variables give a basis for future design work.

The design is somewhat influenced by the end use. However, the principles are the same. It is necessary to stress only the particular characteristic desired in the original design.

Since most inductors now being designed and used in the radio engineering field are for low power applications, as in receiving equipment, the practical data given in this section pertain to such applications. In high-powered transmitting units less thought is given to space and cost, and more thought is given to losses. Most such inductors are designed for single-frequency operation. They may operate at such high power that cooling is necessary. Low-power transmitting inductors are similar to receiving inductors.

CALCULATION OF INDUCTANCE

It is possible to calculate the true (l-f) inductance of some types of air-cored coils with a fair degree of accuracy. *National Bureau of Standards Circular* 74 has long been the most authoritative source of such information. Some of the more useful formulas contained therein are reproduced here.[1] It must be remembered that for r-f inductors the actual apparent inductance differs from the calculated true inductance because of distributed capacitance. For that reason, inductance calculations are generally used only for a starting point in the final design.

In the following formulas all dimensions are expressed in centimeters, and the inductance is in microhenrys.

[1] Rosa, E. B., and F. W. Grover, *Nat. Bur. Standards Sci. Paper* 169; Grover, F. W., *Nat. Bur. Standards Sci. Papers* 320, 1917; 455, 1922; 468, 1923. See for coil design and calculation, especially at low frequencies, Morgan Brooks and H. M. Turner, Inductance of Coils, *Bull.* 53, *Univ. Illinois Eng. Exp. Sta.*, Jan. 8, 1912. Grover, F. W., "Inductance Calculations," Van Nostrand, 1946.

21. Straight Round Wire. If l is the length of the wire, d is the diameter of the cross section, and μ is the permeability of the material of the wire,

$$L_0 = 0.002l\left(\log_\epsilon \frac{4l}{d} - 1 + \frac{\mu}{4}\right) \tag{6}$$

$$= 0.002l\left(2.303 \log_{10} \frac{4l}{d} - 1 + \frac{\mu}{4}\right) \tag{7}$$

If $\mu = 1$ (for all materials except iron),

$$L_0 = 0.002l\left(2.303 \log_{10} \frac{4l}{d} - 0.75\right) \tag{8}$$

The return conductor is assumed to be remote. These formulas give the l-f inductance.

As the frequency increases, the inductance decreases, its value at infinite frequency being

$$L_\infty = 0.002l\left(2.303 \log_{10} \frac{4l}{d} - 1\right) \tag{9}$$

A general expression for the inductance at any frequency is

$$L = 0.002l\left(2.303 \log_{10} \frac{4l}{d} - 1 + \mu\delta\right) \tag{10}$$

The quantity δ is obtained from the table below, as a function of the argument x' where

$$x = 0.1405d\sqrt{\frac{\mu f}{\rho}} \tag{11}$$

and f is the frequency and ρ is the volume resistivity of the wire in microhm-centimeters. For copper at 20°C,

$$x_c = 0.1071d\sqrt{f}$$

This quantity δ will be used in several of the following formulas without further definition.

Value of δ in Inductance Formulas

x	δ	x	δ	x	δ	x	δ	x	δ	x	δ
0	0.250	2.5	0.228	6.0	0.116	12.0	0.059	25.0	0.028	70.0	0.010
0.5	0.250	3.0	0.211	7.0	0.100	14.0	0.050	30.0	0.024	80.0	0.009
1.0	0.249	3.5	0.191	8.0	0.088	16.0	0.044	40.0	0.0175	90.0	0.008
1.5	0.247	4.0	0.1715	9.0	0.078	18.0	0.039	50.0	0.014	100.0	0.007
2.0	0.240	5.0	0.139	10.0	0.070	20.0	0.035	60.0	0.012	∞	0.000

22. Two Parallel Round Wires—Return Circuit. The current is assumed to flow in opposite directions in two parallel wires of length l and diameter d, the distance between centers of wires being D. Then

$$L = 0.004l\left(2.303 \log_{10} \frac{2D}{d} - \frac{D}{l} + \mu\delta\right) \tag{12}$$

This neglects the inductance of the wires connecting the two main wires. If these wires are long, their inductance may be calculated by Eq. (10) and added to the result from Eq. (12), or the whole system may be treated as a rectangle and the inductance calculated by Eq. (14).

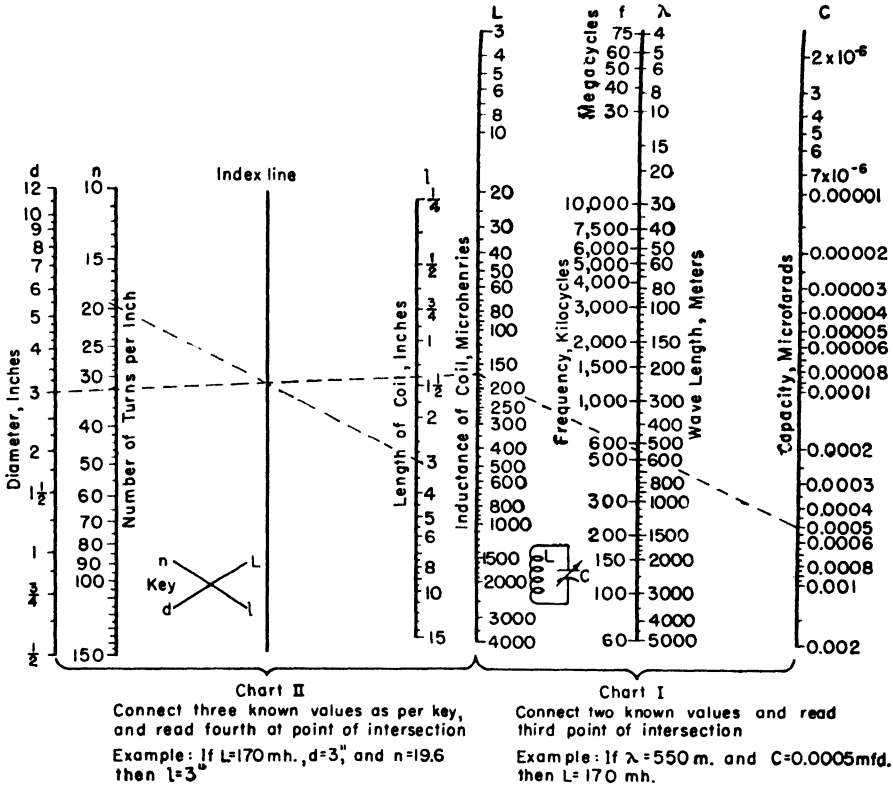



FIG. 10. Inductance design chart.

23. Square of Round Wire. The length of one side of the square is denoted by a; other letters have already been defined.

$$L = 0.008a\left(2.303 \log_{10} \frac{2a}{d} + \frac{d}{2a} - 0.774 + \mu\delta\right) \tag{13}$$

24. Rectangle of Round Wire. The sides of the rectangle are a and a_1 and the diagonal $g = \sqrt{a^2 + a_1^2}$. Then

$$L = 0.00921\left[(a + a_1)\log_{10}\frac{4aa_1}{d} - a\log_{10}(a + g) - a_1\log_{10}(a_1 + g)\right] + 0.004\left[\mu\delta(a + a_1) + 2\left(g + \frac{d}{2}\right) - 2(a + a_1)\right] \tag{14}$$

25. Grounded Horizontal Wire. The wire is assumed to be parallel to the earth' which acts as the return circuit. In addition to symbols already used, h denotes the height of the wire above ground. Then

$$L = 0.004605l\left[\log_{10}\frac{4h}{d} + \log_{10}\left(\frac{l + \sqrt{l^2 + \frac{d^2}{4}}}{l + \sqrt{l^2 + 4h^2}}\right)\right] + 0.002\left(\sqrt{l^2 + 4h^2} - \sqrt{l^2 + \frac{d^2}{4}} + \mu l\delta - 2h + \frac{d}{2}\right) \tag{15}$$

Value of K in Eq. (17)

Diam to length	K	Difference	Diam to length	K	Difference	Diam to length	K	Difference
0.00	1.0000	−0.0209	2.00	0.5255	−0.0118	7.00	0.2584	−0.0047
.05	.9791	203	2.10	.5137	112	7.20	.2537	45
.10	.9588	197	2.20	.5025	107	7.40	.2491	43
.15	.9391	190	2.30	.4918	102	7.60	.2448	42
.20	.9201	185	2.40	.4816	97	7.80	.2406	40
0.25	0.9016	−0.0178	2.50	0.4719	−0.0093	8.00	0.2366	−0.0094
.30	.8838	173	2.60	.4626	89	8.50	.2272	86
.35	.8665	167	2.70	.4537	85	9.00	.2185	79
.40	.8499	162	2.80	.4452	82	9.50	.2106	73
.45	.8337	156	2.90	.4370	78	10.00	.2033	
0.50	0.8181	−0.0150	3.00	0.4292	−0.0075	10.0	0.2033	−0.0133
.55	.8031	146	3.10	.4217	72	11.0	.1903	113
.60	.7885	140	3.20	.4145	70	12.0	.1790	98
.65	.7745	136	3.30	.4075	67	13.0	.1692	87
.70	.7609	131	3.40	.4008	64	14.0	.1605	78
0.75	0.7478	−0.0127	3.50	0.3944	−0.0062	15.0	0.1527	−0.0070
.80	.7351	123	3.60	.3882	60	16.0	.1457	63
.85	.7228	118	3.70	.3822	58	17.0	.1394	58
.90	.7110	115	3.80	.3764	56	18.0	.1336	52
.95	.6995	111	3.90	.3708	54	19.0	.1284	48
1.00	0.6884	−0.0107	4.00	0.3654	−0.0052	20.0	0.1236	−0.0085
1.05	.6777	104	4.10	.3602	51	22.0	.1151	73
1.10	.6673	100	4.20	.3551	49	24.0	.1078	63
1.15	.6573	98	4.30	.3502	47	26.0	.1015	56
1.20	.6475	94	4.40	.3455	46	28.0	.0959	49
1.25	0.6381	−0.0091	4.50	0.3409	−0.0045	30.0	0.0910	−0.0102
1.30	.6290	89	4.60	.3364	43	35.0	.0808	80
1.35	.6201	86	4.70	.3321	42	40.0	.0728	64
1.40	.6115	84	4.80	.3279	41	45.0	.0664	53
1.45	.6031	81	4.90	.3238	40	50.0	.0611	43
1.50	0.5950	−0.0079	5.00	0.3198	−0.0076	60.0	0.0528	−0.0061
1.55	.5871	76	5.20	.3122	72	70.0	.0467	48
1.60	.5795	74	5.40	.3050	69	80.0	.0419	38
1.65	.5721	72	5.60	.2981	65	90.0	.0381	31
1.70	.5649	70	5.80	.2916	62	100.0	.0350	
1.75	0.5579	−0.0068	6.00	0.2854	−0.0059			
1.80	.5511	67	6.20	.2795	56			
1.85	.5444	65	6.40	.2739	54			
1.90	.5379	63	6.60	.2685	52			
1.95	.5316	61	6.80	.2633	49			

26. Circular Ring of Circular Section. If a is the mean radius of the ring,

$$L = 0.01257a\left(2.303\log_{10}\frac{16a}{d} - 2 + \mu\delta\right) \tag{16}$$

provided that $d/2a \leqslant 0.2$.

27. Single-layer Coil or Solenoid.

$$L = \frac{0.0395a^2n^2}{b}K \tag{17}$$

where n is the number of turns, a is the radius of the coil measured from the axis to the center of the wire, b is the length of the coil, and K is a function of $2a/b$, the value of which may be determined by means of the table above.

28. Multilayer Coils: Circular Coils of Rectangular Cross Section. For long coils of a few layers, the following formula may be used:

$$L = L_s - \frac{0.0216n^2ac}{b}(0.693 + B_s) \tag{18}$$

where L_s is the inductance calculated by Eq. (17), n and b are the same as in Eq. (17), a is the radius of coil measured from axis to center of winding cross section, c is the radial depth of winding, and B_s is the correction given below.

Value of B_s in Eq. (18)

b/c	B_s	b/c	B_s	b/c	B_s	b/c	B_s	b/c	B_s	b/c	B_s
1	0.0000	6	0.2446	11	0.2844	16	0.3017	21	0.3116	26	0.3180
2	0.1202	7	0.2563	12	0.2888	17	0.3041	22	0.3131	27	0.3190
3	0.1753	8	0.2656	13	0.2927	18	0.3062	23	0.3145	28	0.3200
4	0.2076	9	0.2730	14	0.2961	19	0.3082	24	0.3157	29	0.3209
5	0.2292	10	0.2792	15	0.2991	20	0.3099	25	0.3169	30	0.3218

For short multilayer coils, the dimensions shown in Fig. 11 are used. Two formulas are required, one for use when $b > c$, and the other for use when $b < c$. In the first case,

$$L = 0.01257an^2\left[\left(1 + \frac{b^2}{32a^2} + \frac{c^2}{96a^2}\right)\log_\epsilon \frac{8a}{d} - y_1 + \frac{b^2}{16a^2}y^2\right]$$

$$= 0.01257an^2\left[2.303\left(1 + \frac{b^2}{32a^2} + \frac{c^2}{96a^2}\right)\log_{10} \frac{8a}{d} - y_1 + \frac{b^2}{16a^2}y^2\right] \tag{19}$$

Value of Constants in Eqs. (19) and (20)

b/c or c/b	y_1	c/b	y_2	b/c	y_3
0	0.5000	0	0.125	0	0.597
0.025	0.5253				
0.05	0.5490	0.05	0.127	0.05	0.599
0.10	0.5924	0.10	0.132	0.10	0.602
0.15	0.6310	0.15	0.142	0.15	0.608
0.20	0.6652	0.20	0.155	0.20	0.615
0.25	0.6953	0.25	0.171	0.25	0.624
0.30	0.7217	0.30	0.192	0.30	0.633
0.35	0.7447	0.35	0.215	0.35	0.643
0.40	0.7645	0.40	0.242	0.40	0.654
0.45	0.7816	0.45	0.273	0.45	0.665
0.50	0.7960	0.50	0.307	0.50	0.677
0.55	0.8081	0.55	0.344	0.55	0.690
0.60	0.8182	0.60	0.384	0.60	0.702
0.65	0.8265	0.65	0.427	0.65	0.715
0.70	0.8331	0.70	0.474	0.70	0.729
0.75	0.8383	0.75	0.523	0.75	0.742
0.80	0.8422	0.80	0.576	0.80	0.756
0.85	0.8451	0.85	0.632	0.85	0.771
0.90	0.8470	0.90	0.690	0.90	0.786
0.95	0.8480	0.95	0.752	0.95	0.801
1.00	0.8483	1.00	0.816	1.00	0.816

When $b < c$,

$$L = 0.01257an^2\left[\left(1 + \frac{b^2}{32a^2} + \frac{c^2}{96a^2}\right)\log_\epsilon \frac{8a}{d} - y_1 + \frac{c^2}{16a^2}y^3\right]$$

$$= 0.01257an^2\left[2.303\left(1 + \frac{b^2}{32a^2} + \frac{c^2}{96a^2}\right)\log_{10} \frac{8a}{d} - y_1 + \frac{c^2}{16a^2}y^3\right] \quad (20)$$

y_1, y_2, and y_3 may be obtained from the table shown on page 3-19. These formulas are quite accurate as long as the diagonal of the cross section (d, Fig. 11) does not exceed the mean radius. The accuracy decreases considerably as b becomes large in comparison with a.

For very accurate results, a correction must be added if the insulation of the wire occupies a considerable percentage of the winding space. This correction is given by

$$\Delta L = 0.01257an\left(2.303 \log_{10} \frac{D}{d} + 0.155\right) \quad (21)$$

where D is the distance between the centers of adjacent wires, and d is the diameter of the bare wire.

Fig. 11. Multilayer coil.

29. Multilayer Square Coil. If n is the number of turns and a is the side of the square measured to the center of the rectangular cross section which has length b and depth c, then

$$L = 0.008an^2\left(2.303 \log_{10} \frac{a}{b+c} + 0.2235 \frac{b+c}{a} + 0.726\right) \quad (22)$$

If the cross section is square ($b = c$), this becomes

$$L = 0.008an^2\left(2.303 \log_{10} \frac{a}{b} + 0.447 \frac{b}{a} + 0.033\right) \quad (23)$$

Equation (21) may be used to correct for insulation by replacing the factor 0.01257 by 0.008. For a single-layer square coil,

$$L = 0.008an^2\left(2.303 \log_{10} \frac{a}{b} + 0.2231 \frac{b}{a} + 0.726\right) - 0.008an(A + B) \quad (24)$$

Value of A in Eq. (24)

d/D	A	d/D	A	d/D	A
1.00	0.557	0.40	−0.359	0.15	−1.340
0.95	0.506	0.38	−0.411	0.14	−1.409
0.90	0.452	0.36	−0.465	0.13	−1.483
0.85	0.394	0.34	−0.522	0.12	−1.563
0.80	0.334	0.32	−0.583	0.11	−1.650
0.75	0.269	0.30	−0.647	0.10	−1.746
0.70	0.200	0.28	−0.716	0.09	−1.851
0.65	0.126	0.26	−0.790	0.08	−1.969
0.60	0.046	0.24	−0.870	0.07	−2.102
0.55	−0.041	0.22	−0.957	0.06	−2.256
0.50	−0.136	0.20	−1.053	0.05	−2.439
0.48	−0.177	0.19	−1.104	0.04	−2.662
0.46	−0.220	0.18	−1.158	0.03	−2.950
0.44	−0.264	0.17	−1.215	0.02	−3.355
0.42	−0.311	0.16	−1.276	0.01	−4.048

Value of B in Eq. (24)

Number of turns, n	B	Number of turns, n	B
1	0.000	40	0.315
2	0.114	45	0.317
3	0.166	50	0.319
4	0.197	60	0.322
5	0.218	70	0.324
6	0.233	80	0.326
7	0.244	90	0.327
8	0.253	100	0.328
9	0.260	150	0.331
10	0.266	200	0.333
15	0.286	300	0.334
20	0.297	400	0.335
25	0.304	500	0.336
30	0.308	700	0.336
35	0.312	1,000	0.336

A and B are given above, where d is the diameter of the bare wire and D is the distance between turns, measured to the centers of the wires.

30. Calculation of Mutual Inductance.[1] The mutual inductance of two parallel coaxial circles may be calculated by the following method: first, calculate

$$\frac{r_2}{r_1} = \sqrt{\frac{\left(1 - \frac{a}{A}\right)^2 + \frac{D^2}{A^2}}{\left(1 + \frac{a}{A}\right)^2 + \frac{D^2}{A^2}}} \tag{25}$$

where a is the radius of the smaller circle, A the radius of the larger circle, and D the distance between the planes of the two circles. From the table on page 3-22 the value of F corresponding to the calculated value of r_2/r_1 is obtained. Then

$$M = F\sqrt{Aa} \tag{26}$$

The units are the same as in the formulas for self-inductance already given.

For two parallel coaxial multilayer coils of square or nearly square cross section, a good approximation is given by

$$M = n_1 n_2 M_0 \tag{27}$$

where n_1 and n_2 are the numbers of turns on the two coils, and M_0 is the mutual inductance of two circles located at the centers of the cross sections of the two coils.

The same formula may be used as a rough approximation for the mutual inductance of two coaxial single-layer solenoids.

References

Butterworth, S.: The High-frequency Copper Losses in Inductance Coils, *Exptl. Wireless*, **2**, 613, 1925.

Coursey, P. R.: Calculation and Design of Inductances, *Electrician*, **75**, 841, 1915.

Griffiths, W. H. F.: Notes on Standard Inductances for Wave Meters and Other Radio-frequency Purposes, *Exptl. Wireless and Wireless Engr.*, **6**, 543, 1929.

High-frequency Resistance and Capacitance of Single-layer Solenoids, *Wireless Engr.*, February and March, 1940.

Kantor, Myron: Theory and Design of Progressive and Ordinary Windings, *Proc. IRE*, **35**, 1563, December, 1947.

Pernot, F. E.: Formulae and Tables for Design of Air-core Inductance Coils, *Univ. Calif. Pub., Eng.*, **1**, 117, 1916.

Soucy, C. I.: Temperature Coefficient Effects of R-f Coil Finishes, *Tele. Tech.*, **6**, 53, December, 1947.

[1] Rosa, E. B., and F. W. Grover, *Nat. Bur. Standards Sci. Paper* 169; Grover, F. W., *Nat. Bur. Standards Sci. Papers* 320 and 498.

Values of F in Eq. (26)

r_2/r_1	F	Difference	r_2/r_1	F	Difference	r_2/r_1	F	Difference
0	∞							
0.010	0.05016	−0.00120	0.30	0.008844	−0.000341	0.80	0.0007345	−0.0000604
.011	4897	109	.31	8503	328	.81	6741	579
.012	4787	100	.32	8175	314	.82	6162	555
			.33	7861	302	.83	5607	531
0.013	4687	−0.00093	.34	7559	290	.84	5076	507
.014	4594	87						
.015	4507	81	0.35	0.007269	−0.000280	0.85	0.0004569	−0.0000484
.016	4426	148	.36	6989	270	.86	4085	460
.018	4278	132	.37	6720	260	.87	3625	437
			.38	6460	249	.88	3188	413
0.020	0.04146	−0.00119	.39	6211	241	.89	2775	389
.022	4027	109						
.024	3918	100	0.40	0.005970	−0.000232	0.90	0.0002386	−0.0000365
.026	3818	93	.41	5738	225	.91	2021	341
.028	3725	86	.42	5514	217	.92	1680	316
			.43	5297	210	.93	1364	290
0.030	3639	−0.00081	.44	5087	202	.94	1074	263
.032	3558	76						
.034	3482	71	0.45	0.004885	−0.000195	0.95	0.00008107	−0.00002351
.036	3411	68	.46	4690	189	.96	5756	2046
.038	3343	64	.47	4501	183	.97	3710	1706
			.48	4318	178	.98	2004	1301
0.040	0.03279	−0.00061	.49	4140	171	.99	703	703
.042	3218	58				1.00	0	
.044	3160	55	0.50	0.003969	−0.000166			
.046	3105	53	.51	3803	160	0.950	0.00008107	0.00000494
.048	3052	51	.52	3643	156	.952	7613	482
			.53	3487	150	.954	7131	470
0.050	0.03001	−0.00226	.54	3337	146	.956	6661	458
.060	2775	191				0.958	6202	446
.070	2584	164	0.55	0.003191	−0.000141			
.080	2420	144	.56	3050	137	0.960	0.00005756	−0.00000436
.090	2276	128	.57	2913	133	.962	5320	421
			.58	2780	128	.964	4899	409
0.100	0.02148	−0.00116	.59	2652	125	.966	4490	397
.11	2032	104				.968	4093	383
.12	1928	96	0.60	0.002527	−0.000120			
.13	1832	89	.61	2407	117	0.970	0.00003710	−0.00000370
.14	1743	82	.62	2290	113	.972	3340	356
			.63	2177	109	.974	2984	341
0.15	0.01661	−0.00075	.64	2068	106	.976	2643	327
.16	1586	71				.978	2316	312
.17	1515	66	0.65	0.001962	−0.000103			
.18	1449	62	.66	1859	99	0.980	0.00002004	−0.00000296
.19	1387	59	.67	1760	96	.982	1708	278
			.68	1664	93	.984	1430	262
0.20	0.01328	−0.00055	.69	1571	90	.986	1168	242
.21	1273	52				.988	926	223
.22	1221	50	0.70	0.001481	−0.000087			
.23	1171	47	.71	1394	84	0.990	0.00000703	−0.00000201
.24	1124	45	.72	1310	81	.992	502	177
			.73	1228	78	.994	326	148
0.25	0.010792	−0.000425	.74	1150	76	.996	177	115
.26	10366	408				.998	62	62
.27	9958	388	0.75	0.0010741	−0.0000731			
.28	9570	371	.76	10010	704			
.29	9199	355	.77	9306	680			
			.78	8626	653			
			.79	7973	628			

Whittemore, L. E., and G. Breit: Inductance, Capacity, and Resistance of Coils at Radio Frequency, *Phys. Rev.*, **14**, 170, 1919.
Wilmotte, R. M.: Parasitic Losses in Inductance Coils at Radio Frequency, *Exptl. Wireless*, **2**, 451 and 477, 1925.

31. Coil Form Relations. From the preceding data it will be seen that for solenoid windings the maximum inductance occurs when the diameter of the winding divided by the length of the winding equals 2.46. For maximum inductance in multilayer solenoids, the winding cross section should be square and the diameter of the form should be about equal to twice the length of a side of that square. For solenoids of the same inductance, the distributed capacitance is approximately proportional to the radius of the form and independent of the number of turns and length of winding.

32. Choice of Wire Size. The choice of wire size for universal windings or progressive universals depends on the desired Q, the permissible outside diameter, and the cost. At the frequencies at which these windings are used, roughly 200 to 2,000 kc stranded wire is desirable to give good Q. In general for a given inductance, the greater the number of strands, the higher the cost and the better the Q. The smaller the strand size, the higher the Q and the cost. The smaller the strand size, the less the outside diameter, a desirable relation with iron cores or where shielding is to be used. The strands should be loosely twisted, about 10 or 12 twists per foot for best results. The effect of wire size on Q, relative cost, and relative outside diameter for a 1-mh winding using some common wire sizes is given in Table 7.

In the region from about 2 to 6 Mc, solenoid windings are used with a wire size of from No. 30 to No. 34. From 6 to 100 Mc, the common wire sizes are from No. 18 to No. 24.

33. Choice of Wire Insulation. There are many insulating materials now in use on wire for fabricating inductors. They fall into two basic groups, the filament or textile type and the coated, dipped, or extruded type. Silk, nylon, du Pont's orlon, celanese, cotton, and glass-fiber filaments are wound around the conductors in either single or double layers to produce the first type. Frequently the conductor has been previously coated with an insulator to prevent short circuits between the turns in case the threads should separate. The purpose of the textile is to facilitate the winding of universal windings or to provide proper spacing between the conductors.

Silk has been very popular but is being replaced by synthetic materials because of its high cost. Cotton gives a very bulky insulation and is used only where considerable conductor spacing is required. The only major r-f coil application is on broadcast-band loop antennas. Orlon winds well but has the disadvantage of poor electrical properties at elevated temperatures which may cause a "runaway" condition resulting in an overheated winding. Glass fiber is difficult to handle and is used only where high temperatures are encountered. Celanese is rather fragile and subject to attack by many of the solvents used in cements and in coil-impregnating materials. It will melt away at soldering temperatures and hence does not have to be removed before soldering.

Nylon-served wire introduces some winding problems since it is both slippery and elastic. However, when properly served and carefully handled, it represents a compromise of properties which explains its current popularity.

Coating materials are oleoresinous varnish, polyvinylformal, nylon, polytetrafluoroethylene, polyurethane, and silicone enamel. Oleoresinous varnish is the traditional plain enamel which has been the most common magnet insulation. Polyvinylformal, better known as "Formvar," was introduced to provide an enamel with greater abrasion resistance. Nylon enamel and, recently, polyurethane enamel, which results from converting isocyanates into resinous compositions, were developed to provide electrical and mechanical properties approaching Formvar but with the added feature of "solderability." These materials melt or decompose at soldering temperatures so that they need not be scraped off before making a soldered joint. Polytetrafluoroethylene, Teflon, either plain or combined with ceramic materials and silicone enamels, was developed for high-temperature applications.

For extremely high temperatures, above 300°F, a completely different approach is being tried. Copper wire is electrocoated with aluminum. The aluminum is then anodized to provide an insulating film.

It should be remembered that the insulation on wire is a necessary evil. It should be omitted wherever possible, *e.g.*, in spaced solenoids wound on threaded forms. In any case it should be kept to the practical minimum.

34. Choice of Form. Form size is dictated by space limitations, winding type, winding design, and cost. The important factor left to consider is material. It must be strong enough to support the winding and, since it is usually largely within the field of the winding, it must be electrically good. It should be dimensionally stable with temperature changes. No material yet known is electrically perfect and, therefore, the less form material used, the less additional losses will be introduced.

Common materials are ceramics, such as steatite, thermosetting plastics, such as molded or laminated phenolics, thermoplastics, such as polystyrene and its copolymers, cellulose acetate and its copolymers, ethyl cellulose, glass, fiber, and paper.

Ceramics and glass are good electrically but are fragile and cannot be made to close dimensional tolerances. As a group they are the most expensive. The phenolics vary widely in their electrical properties depending on the filler used. The better electrical grades are satisfactory and are mechanically better and less costly than ceramics. Certain of the plastics, particularly the styrenes and ethyl cellulose, have the best electrical properties but are not too widely used because they will not stand the rather high temperatures encountered in modern receivers. This is particularly true when the winding has appreciable tension.

Fiber and paper are economical. When properly treated with wax, these materials are suitable for all but the most exacting applications.

Certain windings are completely self-supporting and require no form. Others are wound directly on powdered-iron cores.

35. Choice of Finish. The finish on an inductor performs two major functions: it helps to make the unit mechanically stable and to make it electrically stable through protection against ambient conditions. The most common finishes are waxes, varnishes, and lacquers, or synthetics. The choice of finish depends on the conditions to be encountered and on the amount of additional losses that can be tolerated in the final design. The use of any finish always introduces some loss.

Operating temperatures may range from −20 to +160°F for civilian broadcast units. The rigorous demands of high-performance aircraft, guided missiles, and satellites extend the range from −70 to +300, 400, and even 1000°F. At these temperatures the units may also be subjected to high values of shock and vibration. Relative humidities may vary over a wide range.

Waxes, particularly the microcrystalline type, ceresins, etc., are excellent for the moderate temperatures of civilian use. They are simple to use, are inexpensive, have fairly low losses, and low dielectric constants, are noncorrosive, and provide the best humidity protection. They should never be used when the ambient temperature is close to their melting point.

Varnishes, both the solvent and solventless types, are usually not so good electrically, particularly at the higher frequencies. The initial expense and cost of application are high. The common heat polymerizing varnishes have a short shelf life before use. However, properly applied varnishes give good humidity protection, withstand a greater temperature range than waxes, and are mechanically strong.

Lacquers or synthetics are easier to apply than varnishes, give about as good mechanical strength, have lower losses, but do not give so good humidity protection. They can be made so that they have low dielectric constants. This fact is important since any finish becomes a part of the dielectric of the distributed capacitance; in many designs it is a large part.

In choosing a finish the following points are important: (1) the temperature range. The finish must hold the windings firmly throughout the operating temperature range. It must not change physically with temperature changes for that would distort the

winding. It must not change its dielectric constant since that would change the apparent inductance. (2) The finish should be as good a dielectric as possible so that the losses will be kept to a minimum. The dielectric constant must be low to keep the distributed capacitance low. (3) The finish should afford good humidity protection and at the same time be noncorrosive. Since many inductors have d-c potentials applied and are wound with fine wire or stranded wire, any tendency for corrosion would be accelerated and quickly result in an open circuit.

Accelerated corrosion tests can be made with typical universal windings spaced ¼ in. apart, wound with 7/41 SSE litz wire, subjected to 90 to 95 per cent relative humidity at 100 to 120°F, with 400 volts d.c. between the windings. Under these conditions, windings treated with practically any lacquer or varnish show at least one and at most seven strands of the litz completely corroded after 96 hr. Wax-treated units in the same test rarely show more than one broken strand and usually not any.

36. Methods of Obtaining Close Coupling. For universal windings, close coupling is obtained by making the second winding a continuation of the first, sometimes separated by a layer of tape. The two windings may be placed side by side very close together. In some cases maximum coupling has been obtained by bifilar winding in which two conductors are wound simultaneously and are always parallel as the winding is built up.

For solenoids bifilar winding is also used. Less coupling is obtained by interwinding only a part of the two windings. The second winding may be placed directly over the first, separated by a tape.

Similar methods apply to other types of windings. Where powdered-iron cores are used, close coupling can be obtained by simply winding both coils on the same iron. Such practice is not recommended unless high-resistance iron is used.

Coupling as high as 98 per cent can be obtained with bifilar universals and as high as 90 per cent with bifilar solenoids. The use of iron cores in the windings and iron cups surrounding the windings will increase the coupling in most cases.

37. Effect of Shielding. To prevent undesirable coupling between inductors it is common to surround them with a shield of conducting material. The best conductor is copper and, although it is sometimes used, it is expensive and requires some treatment to maintain its appearance. Easy to fabricate, light in weight, low in cost, and requiring no treatment, aluminum is a favorite shield material. Zinc is also used, especially when it is necessary to soft-solder to the shield.

Shielding is accomplished because of the eddy currents set up in the conducting material. The eddy currents oppose the inducing flux and, therefore, repel and confine it. This opposition to the flux acts as a negative permeability of the surrounding medium. Because the shield is not a perfect conductor, losses are introduced that reflect as losses in the inductor. Shielding lowers the inductance and the Q. The proximity of the shield to the inductor and the shield material controls the amount of lowering that takes place.

Shielding is also accomplished through the use of cups or sleeves of powdered iron. In this case the flux is concentrated in the low-reluctance path of the iron. The Q and inductance of the winding may be slightly increased. This type of shielding is not complete, and usually the whole is surrounded with a conventional aluminum shield, which then has little effect on the winding. Its function is to provide electrostatic shielding for more complete shielding.

38. Special Considerations in the Use of Powdered-iron Cores. The use of powdered-iron cores materially changes some details of winding design. It is most important to get as much of the winding as possible as near the core as possible. In so doing the maximum number of turns of the conductor is affected by the core. In other words, there is less leakage flux and more active flux to produce the maximum inductance.

In the case of universal windings, the depth of the winding should be kept to a minimum to keep the average turn closer to the core. In all cases, the less the wall thickness of the form used, the better the results.

39. Typical Inductors Currently in Use. The following inductors are used in a typical broadcast receiver employing a loop antenna. The set has a standard 365-$\mu\mu$f three-gang variable capacitor. All inductors are impregnated and flashed in wax. Shields used for r-f coil, first and second i-f coils, are all 1¼-in.-square aluminum, 2½ in. long.

Loop Antenna.

Wound in the form of a flat spiral, elliptical in shape, minor diameter of average turn = 5 in., major diameter of average turn = 11 in., wire = 22 DCE, Q at 1 Mc = 149 far from chassis and 72 mounted on chassis, turns = 27, inductance = 265.7 μh, C_d = 16.8 $\mu\mu$f.

R-f Coil.

Type	Turns	Width, in.	Wire	Air inductance	Q in shield at 1 Mc
Primary universal	646	13/64	39 SSE	6.38 mh	
Secondary 4-layer bank	195	....	7/41 SSE	280 μh	98

Spacing primary to secondary 5/64 in. L_M = 200 μh. Form: ½ in. OD by 7/16 in. ID, bakelite.

Oscillator Coil.

Type	Turns	Width, in.	Wire	Air inductance, μh
Primary universal	23¾	3/16	36 SCE	14
Secondary universal	83¼	3/16	36 SCE	127

Secondary wound on top of primary separated by yellow cambric ⅜ by 0.010 by 211/16 in. L_M = 25.1 μh in air. Form: ½ in. OD by ⅜ in. ID, wax-impregnated paper.

First I-f Transformer.

Type	Turns	Width, in.	Wire	Air inductance, mh	Q in shield at 455 kc
Primary universal	422	3/16	5/42 SSE	2.27	83
Secondary universal	422	3/16	5/42 SSE	2.27	82

Spacing primary to secondary 9/16 in. L_M = 109.2 μh in air. Form: ⅜ in. OD by 5/16 in. ID, wax-impregnated paper.

Second I-f (Diode) Transformer.

Type	Turns	Width, in.	Wire	Air inductance, mh	Q in shield at 455 kc
Primary universal	365	3/16	5/42 SSE	1.7	90
Secondary universal	415	3/16	5/42 SSE	2.2	83

Spacing, primary to secondary, 39/64 in. L_M = 130.7 μh in air. Form: ⅜ in. OD by 5/16 in. ID, wax-impregnated paper.

The following inductors are used in a typical permeability-tuned automobile broadcast receiver covering 535 to 1,610 kc.

Antenna and R-f Coils.

Form: 0.223 in. OD by 0.207 in. ID, bakelite. 356 turns 5/43 SSE straight progressive universal winding 1 1/4 in. long. Minimum Q = 60 at 600 kc. Core: 0.200 in. OD by 1 1/2 in. long. Apparent permeability = 9. Lacquer-coated. These coils tune to 1,610 kc with 65 μμf, no core.

Oscillator Coil.

Form: 0.318 in. OD by 0.207 in. ID, bakelite. 140 turns 36 E. Variable-pitch solenoid, 1 9/32 in. long. Core same as for antenna and r-f coils. Lacquer-coated. This coil tunes to 1,870 kc with 196 μμf, no core. Cathode coil wound on 7/16 in. OD by 3/8 in. ID bakelized paper. Winding: 16 turns of No. 30 E at 13 turns per inch. This coil is slipped over the oscillator winding and exactly centered over it.

The i-f transformers used in this receiver are listed as "auto set" transformers in Table 13. The transformers in this table all have the same basic construction. All have 3/4-in.-square aluminum shields, 0.283-in.-OD phenolic-impregnated tubing, threaded to receive 1/4 in. by 28 pitch, powdered-iron screw cores and, where required, silvered-mica capacitors.

Table 13. Typical Receiver Inductors

Use	Winding	Frequency	Type of winding	C/T or Tpi	Turns	Wire	Core	Inductance, mh		Mutual inductance, mh or spacing (sp.)	Q	Capacity, μμf	Width, in.	OD, in.
								Max	Peaked					
Auto set Input I.F.	Pri.	1	1	1 3/5	518	1	1	3.55	2.56	143	44	131	0.150	0.367
	Sec.	1	1	1 3/5	647	1	1	5.35	3.73	.....	46	86	0.150	0.415
Auto set Output I.F.	Pri.	1	1	1 3/5	522	4	4	3.44	2.72	132	74	125	0.200	0.400
	Sec.	1	1	1 3/5	588	5	4	4.24	3.03	....	44	125	0.200	0.552
TV Pix I.F.	Pri.	3	2	100	12	3	5	Wound to turns		0.117	98			
	Sec.	3	2	100	17	3	5	....		sp.	99			
TV sound Trap	....	4	2	CW	11½	8	6	Wound to turns		.....	135	18	.....	0.340
TV sound Pickup	....	2	1	3	44	2	3	0.041	0.031	.....	47	47	0.062	0.365

NOTES:
Frequency: 1–262 kc, 2–4.5 Mc, 3–44 Mc, 4–41.25 Mc.
Type of winding: 1, universal; 2, solenoid.
Core: 1, Plast-iron B214; 2, Carbonyl TH; 3, Carbonyl L; 4, Carbonyl C; 5, Carbonyl J; 6, IRN8.
Wire: 1, No. 40 SNE; 2, No. 38 SNE; 3, No. 36 SNE; 4, No. 2/41 SNE; 5, No. 36 E; 6, No. 34 E; 7, No. 30 E; 8, No. 22 E.
Tpi is turns per inch for solenoid windings.
C/T is crosses per turn for universal windings.
CW is close-wound.

Part 2: MAGNETIC MATERIALS[1]

Magnetic materials are essential in the construction of nearly all types of electrical machinery and apparatus. In the field of radio communication, they are employed in the cores of transformers, choke coils, relays, etc., and for magnetic shielding. Permanent magnets are used in d-c measuring instruments, polarized relays, magnetrons, and loud-speakers. A great variety of magnetic materials is produced commercially, and new types are continually being developed. In view of the diversity in requirements for practical applications, a knowledge of the magnetic characteristics of available materials is needed in connection with the design of various types of appa-

[1] The material for this portion of Chap. 3 came in part from "Permanent Magnets" and "Magnetic Testing," *Nat. Bur. Standards Circ.*, C448 and C456.

ratus. The principal producers of magnetic materials issue technical data giving the properties of the materials they make. Before undertaking the design of apparatus involving the use of magnetic materials it is best to obtain up-to-date information from one or more of these manufacturers.

MAGNETIC QUANTITIES AND UNITS

The characteristics of magnetic circuits and the magnetic materials of which they are made are expressed in terms of certain magnetic quantities and units which may be described as follows:[1]

40. Magnetomotive Force. Magnetization is generally accomplished by means of electric current in windings linked with a magnetic circuit of which the material to be magnetized forms a part. The total measure of the magnetizing effect of such a coil is called the magnetomotive force, $\mathfrak{F}$. The cgs unit of magnetomotive force (mmf) is called the *gilbert* and is defined by the equation

$$\mathfrak{F} = 0.4\pi NI$$

where $\mathfrak{F}$ = mmf, gilberts
N = number of turns in the coil
I = current, amp

41. Magnetic Flux. The total measure of the magnetized condition of a magnetic circuit when acted upon by a magnetomotive force is called the "magnetic flux," ϕ. It is characterized by the fact that a variation in its magnitude gives rise to an emf in an electric circuit linked with it. The emf thus induced is at any instant directly proportional to the time rate of variation of the flux. The cgs unit of magnetic flux is called the *maxwell* and is defined by the equation

$$e = N\left(\frac{d\phi}{dt}\right)10^{-8}$$

where e = induced emf, volts
N = number of turns linked with the flux
ϕ = magnetic flux, maxwells
t = time, sec

42. Magnetic Reluctance. That property of a magnetic circuit which determines the relationship between the magnetic flux and the corresponding mmf is called the "magnetic reluctance," $\mathcal{R}$, of the circuit. The cgs unit[2] is defined by the equation

$$\phi = \frac{\mathfrak{F}}{\mathcal{R}}$$

where ϕ = magnetic flux, maxwells
$\mathfrak{F}$ = mmf, gilberts
$\mathcal{R}$ = magnetic reluctance in cgs units

In a magnetic circuit of uniform cross-sectional area and uniform permeability (to be defined later),

$$\mathcal{R} = \frac{l}{\mu A}$$

where $\mathcal{R}$ = magnetic reluctance in cgs units
μ = magnetic permeability
l = length, cm
A = cross-sectional area, sq cm

[1] Although some of these definitions are not mathematically rigorous, they are essentially correct and somewhat easier to understand than the more rigorous forms would be.

[2] The unit of magnetic reluctance was called the "oersted" until 1930, when the International Electrotechnical Commission adopted that name for the unit of magnetizing force, thus leaving the unit of reluctance without a name.

43. Magnetic Permeance. Magnetic permeance, $\mathcal{P}$, is the reciprocal of the magnetic reluctance. Thus,

$$\phi = \mathcal{F}\mathcal{P}$$

where ϕ = magnetic flux, maxwells
$\mathcal{F}$ = mmf, gilberts
$\mathcal{P}$ = magnetic permeance in cgs units

In a circuit of uniform cross section and permeability

$$\mathcal{P} = \frac{\mu A}{l}$$

44. Magnetizing Force. The mmf acting on a magnetic circuit is distributed along its length in a manner determined by the distribution of the magnetizing winding and of the reluctance of the circuit. The mmf per unit length along the circuit is called the "magnetizing force," H. The cgs unit is called the *oersted* and is defined by the equation

$$H = \frac{d\mathcal{F}}{dl}$$

where H = magnetizing force, oersteds
$\mathcal{F}$ = mmf, gilberts
l = length, cm

At the center of a very long, uniformly wound solenoid having n turns per cm in which there is a current of I amp, the magnetizing force in oersteds is

$$H = 0.4\pi nI$$

45. Magnetic Induction. Magnetic induction, B, also called "magnetic flux density," is the magnetic flux per unit area of a section normal to the direction of the flux. The cgs unit is called the *gauss* and is defined by the equation

$$B = \frac{d\phi}{dA}$$

where B = magnetic induction, gausses
ϕ = magnetic flux, maxwells
A = area, sq cm

46. Intrinsic Induction. That part of the magnetic induction which is in excess of the induction that would exist in a vacuum under the influence of a given magnetizing force is called the "intrinsic induction," B_i. Numerically, in the cgs system $B_i = B - H$.

47. Nature of Magnetism. It is generally conceded that the only known source of magnetic effects is electricity in motion (electric current). According to present theory,[1] ferromagnetic effects are due to groups of electrons within a ferromagnetic material called "domains" and consisting of electrons spinning on their own axes. The magnetic axes of the spinning electrons in a single domain are held parallel to each other by mutual forces known as "exchange forces," so that each domain behaves as a single unit. The domains are in effect current-turns and so account for the mmfs inherent in ferromagnetic materials. The sum total of the mmfs due to the domains is the quantity that remains "permanent" in a permanent magnet.

In the unmagnetized condition, the domains are so oriented with respect to each other that the net magnetic effect is zero in any direction. Under the influence of a magnetic field applied by means of an external electric current, the magnetic axes of the domains tend to be oriented more or less in the direction of the applied field, so that their effect is added to that of the applied field. The flux density due to the combined effect of the applied field and the domains is the magnetic induction, B, and that due to the domains alone is the intrinsic induction, B_i.

[1] Bozorth, R. M., "Ferromagnetism," p. 423, Van Nostrand, 1951.

Upon removal of the magnetizing force, the orientations of the domains resulting from the applied field tend to persist to a greater or less extent, depending upon their distribution in space, which in turn depends upon the internal structure of the material. To reduce the induction further, it is necessary to apply a reverse, or demagnetizing, force. The relationships between induction and the applied field are shown graphically in Fig. 12, where the abscissas represent values of magnetizing force, H, and the ordinates are values of magnetic induction, B. The curve *oabc* is called the "normal induction curve" and shows how the induction varies as the magnetizing force applied to a material initially in the demagnetized condition is increased from zero. At any point on this curve, the ratio of the value of induction to the corresponding value of magnetizing force is called the *magnetic permeability*, μ. Thus the coordinates of point a on the curve are B_a and H_a, respectively, and the permeability is

$$\mu = \frac{B_a}{H_a}$$

Fig. 12. Normal induction and hysteresis curves.

Since the line *oabc* is not straight, the permeability is not constant but varies with the degree of magnetization. This variation is one of the distinguishing characteristics of ferromagnetic materials.

48. Hysteresis Effects. If, after the magnetizing force has reached a certain value (as at point b), it is then decreased the induction does not follow the curve *oabc* in reverse order but lags behind the magnetizing force, as shown by the curve *bd*. Thus, when the magnetizing force has been reduced to zero, point d, the induction still has an appreciable value. This value is called the *residual induction*, B_r.

To reduce the induction still further, it is necessary to increase the magnetizing force in the opposite direction. The value of the reversed magnetizing force required to reduce the induction to zero is called the *coercive force*, H_c.

The lagging of the induction behind the magnetizing force is called *hysteresis*, and the complete curve *bdefghb*, is called a hysteresis "loop." Hysteresis is another of the characteristics of ferromagnetic materials and is the property that makes permanent magnets possible. The size of the hysteresis loop, and consequently the magnitudes of B_r and H_c, depends upon the values of B and H at the tip of the loop, generally designated by the symbols B_m and H_m. As H_m is increased, the size of the hysteresis loop approaches a maximum. The values of B_r and H_c corresponding to the maximum loop for a given material are called the *retentivity* and *coercivity*, respectively.

A major hysteresis loop is produced by varying the magnetizing force continuously from a positive maximum value to the same value negative and back again to the positive maximum. However, if a change in direction of the variation is made at some intermediate point on the major loop, a minor loop is produced, as shown between the points i and k in Fig. 12. The slope of the line joining the tips of such minor loops is called the *incremental permeability*, μ_Δ. In other words, the incremental permeability is the ratio of the change in B to the corresponding change in H when the average value of B differs from zero. That is, $\mu_\Delta = \Delta B/\Delta H$.

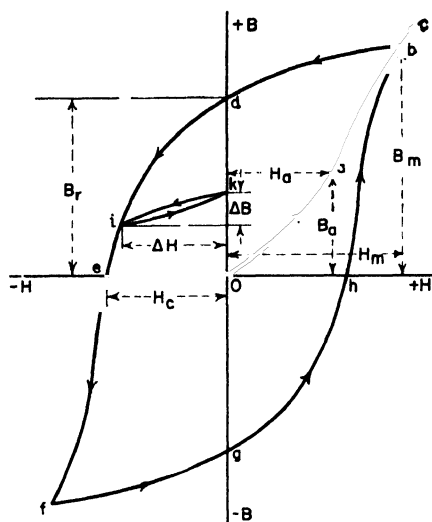
That part of the hysteresis loop extending from the residual induction, B_r, to the coercive force, H_c, *i.e.*, from d to e, is called the *demagnetization curve*. Points on this

curve are designated by the coordinates B_d and H_d. Most of the important characteristics of a permanent-magnet material can be indicated by points on this curve or within the area between it and the axes of coordinates. The value of H_m necessary to produce the maximum loop depends upon the type of material and ranges from about 300 oersteds for most of the older types of material to about 3,000 oersteds or more for the newer types. Figure 13 represents a typical demagnetization curve, together with the so-called *energy-product* curve obtained by plotting the product of corresponding values of B_d and H_d on the demagnetization curve against B_d. The maximum value of this product, $(B_dH_d)_{max}$, can be considered as a criterion of quality and is useful for the comparison of various materials. It is generally not necessary to plot the B_dH_d curve to determine the maximum energy product. A line drawn from the origin of coordinates to the intersection of horizontal and vertical lines through B_r and H_c, respectively, intersects the demagnetization curve at the point where the product B_dH_d is a maximum for most materials.

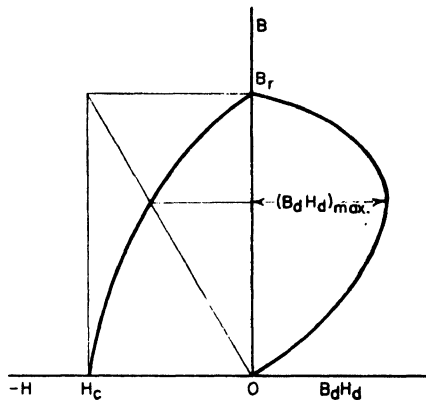
FIG. 13. Demagnetization and energy-product curves.

MAGNETIC CHARACTERISTICS OF MATERIALS

Magnetic materials employed in commercial practice may be considered under the following classifications: (1) solid-core materials, (2) electrical sheet and strip, (3) special-purpose alloys, (4) permanent-magnet alloys.

49. Solid-core Materials. These materials are used for the cores of d-c electromagnets, relays, field frames of d-c machines, etc. The principal requirement is high permeability, particularly at relatively high induction. For most uses it is also desirable that the hysteresis be low. The principal materials employed are soft iron, relay steel, cast steel, cast iron, and an alloy of approximately 35 per cent of cobalt and 65 per cent of iron known as ferrocobalt. Ferrocobalt is characterized by very high permeability in the upper part of the normal induction range and a saturation induction approximately 10 per cent greater than that of pure iron. Its cost is relatively high, however, and its use is limited in general to pole pieces in which a very high induction is required. Cast iron has a relatively low permeability and is used principally in field frames when cost is of primary importance and extra weight is not objectionable. Several varieties of soft iron are available, such as Norway iron, Armco iron, and Swedish charcoal iron. Relay steels contain from 1 to 3.25 per cent of silicon to reduce aging. Electrolytic iron may also be used. All these materials are relatively pure iron, low in carbon and other impurities. Cast steel should be low in carbon, not over 0.1 to 0.2 per cent, and contain only the usual small amounts of the ordinary impurities. Cast iron is high in carbon, about 3 per cent, and also contains about 3 per cent of silicon and varying percentages of phosphorus, manganese, and sulfur.

The best magnetic properties are obtained by a suitable annealing treatment. The properties of cast iron can be greatly improved by malleableizing, a process that converts a large part of the carbon to the amorphous form.

50. Electrical Sheet and Strip. The terms *electrical sheet* and *electrical strip* are commonly used to designate silicon-iron alloys produced in sheet or strip form and used as core materials in a-c apparatus, such as transformers, generators, motors, electromagnets, or relays. The principal requirements are high permeability, low hysteresis, and high resistivity. The several grades differ mainly with respect to their silicon content, which ranges from about 0.5 per cent to approximately 4.5 per

cent. Alloys containing the higher percentages of silicon are practically nonaging; *i.e.*, the permeability and losses do not change with time. The required magnetic properties are produced by annealing.

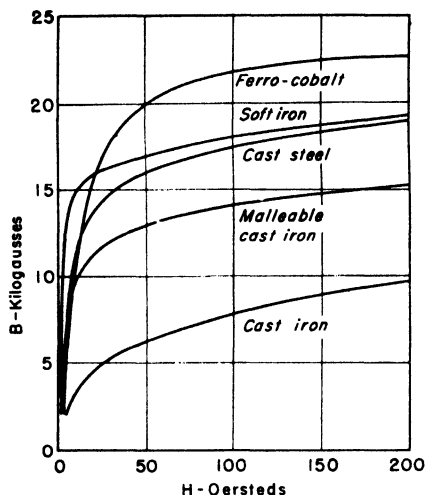



FIG. 14. Typical normal induction curves for solid-core materials.

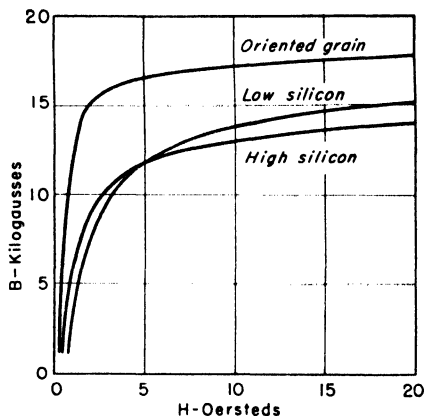



FIG. 15. Typical normal induction curves for electrical sheet.

By a suitable combination of cold-rolling and heat-treatment, materials are produced in which the crystal axes are given a definite orientation. Such material has considerably better properties, when magnetized in the preferred direction, than the ordinary grades. Figure 15 shows typical normal induction curves for two grades of

Table 14. Typical Characteristics of Commercial Electrical Sheet*

Grade	Si per cent	Core loss†		Max permeability		Coercive force, H_c	Resistivity, microhm-cm	Density, g/cm³
		B_m = 10,000	B_m = 15,000	μ_m	B at μ_m			
Armature	0.5	1.30	3.06	5,800	7,000	0.70	19	7.83
Electrical	1.0	1.17	2.75	6,150	6,000	0.68	26	7.79
Motor	2.5	1.01	2.37	6,100	6,100	0.58	42	7.70
Dynamo	3.25	0.82	1.93	5,800	5,000	0.50	50	7.65
Transformer 1	4.0	0.72	1.69	6,300	5,000	0.40	56	7.60
Transformer 2	4.5	0.58	1.36	8,300	4,500	0.25	62	7.57
Oriented grain‡	3.3	0.04	0.09	40,000	6,000	0.10	48	7.65

* Subject to some variation in actual practice.
† Watts per pound at 60 cycles.
‡ Flux in direction of rolling.

electrical sheet and oriented-grain material. The improvement in the oriented-grain material is particularly conspicuous in the upper part of the normal induction curve.

The different grades of electrical sheet and strip are usually sold on the basis of guaranteed maximum values of total core loss, as determined in accordance with the specifications of the ASTM.[1] The common designations of the various grades

[1] "Ferrous Metals," latest ed., ASTM Standards.

are armature, electrical, motor, dynamo, and transformer. The transformer grades are further subdivided into classes denoted by numerals corresponding to the core loss under standard conditions. Armature, electrical, and motor grades are used principally in small motors, a-c magnets, and starting transformers. The dynamo grade is used in high-efficiency rotating machines and small transformers. The transformer grades are used in power and radio transformers (see Table 14).

51. Special-purpose Alloys. For certain applications, special alloys have been developed which, after proper heat-treatment, have superior properties in certain ranges of magnetization. For instance, alloys of nickel and iron with possible small percentages of molybdenum or chromium have very high values of initial and maximum permeability. Alloys of this class, which may have from 70 to 80 per cent of nickel, are called "permalloys." An alloy of 50 per cent of nickel and 50 per cent of iron is called "hipernik." Another alloy having a small percentage of copper in its composition is called "mumetal." The characteristics of these alloys differ in detail, but in general they have high initial and maximum permeability, low hysteresis, and low saturation values. They are particularly applicable for use at low inductions. Typical permeability curves are given in Fig. 16.

Fig. 16. Typical permeability curves for special-purpose alloys.

A certain alloy of nickel, cobalt, and iron after suitable heat-treatment has very nearly constant permeability for inductions below 1,000 gausses and is called *perminvar*. The 50–50 nickel-iron alloy can also be heat-treated so as to have similar characteristics.

An alloy of equal proportions of iron and cobalt has high permeability which persists at higher values of induction than the nickel-iron alloys and is called *permendur*.

Table 15. Typical Characteristics of Some Special Magnetic Alloys*

Alloy	Nominal composition, per cent†	Initial permeability, μ_0	Max permeability, μ_m	Coercive force, H_c	Saturation induction, B_s	Resistivity, microhm-cm	Density, g/cm³
Supermalloy	79Ni, 5Mo	100,000	800,000	0.004	8,000	60	8.87
78 permalloy	78Ni	8,000	100,000	0.05	10,000	16	8.60
4-79 permalloy	79Ni, 4Mo	20,000	80,000	0.05	8,700	57	8.72
45 permalloy	45Ni	2,500	25,000	0.3	16,000	50	8.17
Hipernik	50Ni	4,000	80,000	0.05	16,000	35	8.25
4750 alloy	47 to 50Ni	5,000	32,000	0.03	16,000	45	8.15
1040 alloy	72Ni, 14Cu, 3Mo	40,000	100,000	0.02	6,000	56	8.76
Mumetal	75Ni, 2Cr, 5Cu	20,000	110,000	0.02	7,200	42	8.60
Permendur	50Co	800	5,000	2.0	24,500	7	8.30
Powdered permalloy	81Ni, 2Mo	125	130	.....	......	10^6	7.8

* Subject to some variation in actual practice.
† Balance iron plus usual small amounts of impurities.

Typical characteristics of some special magnetic alloys are given in Table 15.

In recent years, the increasing use of apparatus such as saturable-core reactors, magnetic amplifiers, and the like, which for best operation require materials having hysteresis loops as nearly rectangular as possible, has brought about the development

of a new series of materials.[1] Although the oriented-grain materials previously mentioned have this characteristic to a certain degree, the iron-nickel alloys, when given certain mechanical, thermal, and sometimes magnetic treatments are found to have superior qualities. Not only do they have nearly rectangular hysteresis loops but also their initial and maximum permeabilities are considerably higher.

One caution should be suggested at this point. Data given by producers in published tables generally relate to the static properties of the materials. However, it has been found that the so-called "dynamic" hysteresis loops obtained at frequencies even as low as 60 cps are considerably wider than the static loops. It is always best to check with the producer on this point before deciding on a final design. This is a good general rule because magnetic materials are undergoing continuous development and new and superior materials for a particular purpose may become available at any time.

Another series of magnetic alloys of copper, nickel, and iron is temperature-sensitive, having an approximately linear relation between permeability and temperature. These are called "thermalloys." The principal use is in the compensation of watt-hour meters for temperature variations. They are also used in certain types of thermal relays.

Table 16. Typical Characteristics of Some Commercially Available Ferrites

Nominal composition, %	Saturation induction B_s	Initial permeability μ_0 at 1 Mc	Max permeability μ_m	tan δ_m*	Curie point, °C
10NiO, 23ZnO	1,800	1,000	4,300	0.0022	70
10NiO, 24ZnO	2,400	960	2,700	0.0024	100
16NiO, 17ZnO	3,600	520	1,480	0.0009	140
16NiO, 17ZnO	3,100	330	730	0.0009	260
9NiO, 18ZnO, 1MgO	3,240	430	1,220	0.0007	170
9NiO, 14ZnO, 1MnO	3,200	180	600	0.0006	
11NiO, 20ZnO, 1MnO, 2CuO	3,400	380	3,000	0.0006	160
10ZnO, 19MnO	3,500	690	2,000	0.0017	
12ZnO, 17MnO	3,600	770	2,000	0.0016	
13ZnO, 17MnO, 1SiO₂	3,300	870	1,840	0.0022	

* $\tan \delta_m = 1/Q = R/\omega L = \omega CR$.

52. Ferrites. The principal disadvantage of the sheet and strip materials for high-frequency applications is their relatively low electrical resistivity. In recent years, a new type of material having very high resistivity has been developed. Several materials of this type are commercially available which not only have high resistivity but also have usable magnetic properties. They are ceramiclike materials whose composition may be represented by the formula $xMO \cdot Fe_2O_3 + yM'O \cdot Fe_2O_3$, where M and M' are bivalent metals and $x + y = 1$. The metals most commonly used are nickel, zinc, and manganese. Small amounts of copper, magnesium, or silicon may also be present. These materials are usually called ferrites. Their structure is of the type called spinel. Harvey, Hegyi, and Leverenz[2] have coined the name "ferro-spinel," which seems to be more appropriate because the term "ferrite" is well established as denoting practically pure iron when it is one of the constituents found in ferrous alloys.

The process of manufacture involves mixing the powdered components (oxides) in definite proportions, compressing the mixture into the desired shape, and sintering (not melting) at a high temperature. The mixed components react and diffuse to form the desired spinel structure.

[1] Trade names by which these materials are designated include Deltamax, Permanorm, Permeron, Orthonol, and Orthonik.

[2] Harvey, R. L., I. J. Hegyi, and H. W. Leverenz, Ferromagnetic Spinels for Radio Frequencies, *RCA Rev.*, **11**, 321, 1950.

Table 16 gives some typical characteristics of a few commercially available ferrites. However, materials of this type are undergoing continuous development and much improved properties can confidently be expected. It will be noted that the Curie point for these materials is low. This limits the temperatures at which they can be operated.

53. Permanent-magnet Alloys. For permanent magnets, high values of residual induction and coercive force are required. Ordinary high-carbon steels and some special tool steels make fair magnets when properly hardened, but better results are obtained by the use of alloys specially made for the purpose. Development along this line has been rapid during the past few years and is still going on. There are at present three general classes of permanent-magnet materials: those in which the principal alloying elements are tungsten or chromium or both; those in which cobalt is present in substantial amounts; and those in which aluminum and nickel play an important part.

Tungsten, chromium, or cobalt magnet steels are hardened by quenching. The aluminum-nickel alloys are among the so-called dispersion-hardening alloys, the final

Table 17. Typical Characteristics of Several Permanent-magnet Alloys[a]

Type	Nominal composition[b]	Residual induction, B_r	Coercive force, H_c	$(B_dH_d) \times 10^{-6}$	Fabrication[c]
Tungsten	5.5W, 0.6C	10,000	65	0.27	Hot-forge or cast, machine
3½% Cr	3.5Cr, 0.3Mn, 0.9C	9,500	63	0.25	Hot-forge or cast, machine
Low Co	8.5Co, 5Cr, 1W, 0.9C	7,500	120	0.38	Hot-forge or cast, machine
Co-Cr	16Co, 9Cr, 1C	8,000	180	0.61	Hot-forge or cast, machine
High Co	36Co, 5W, 2Cr, 0.8C	9,000	230	0.87	Hot-forge or cast, machine
Alnico 1	12Al, 20Ni, 5Co	7,300	440	1.40	Cast, grind
Alnico 2	10Al, 17Ni, 12.5Co, 6Cu	7,350	560	1.60	Cast or sinter, grind
Alnico 3	12Al, 25Ni	6,900	475	1.38	Cast, grind
Alnico 4	12Al, 28Ni, 5Co	5,300	730	1.30	Cast, grind
Alnico 5	8Al, 14Ni, 24Co, 3Cu	12,500	640	5.00	Cast, grind[d]
Alnico 12	6Al, 18Ni, 35Co, 8Ti	5,700	900	1.80	Cast, grind
Comol	12Co, 17Mo	10,500	250	1.10	Cast or hot-roll and punch
Cunife 1	60Cu, 20Ni	5,800	600	1.96	Cold-roll, machine[e]
Cunife 2	50Cu, 20Ni, 2.5Co	7,300	260	0.78	Cold-roll, machine[e]
Cunico	50Cu, 21Ni, 29Co	3,400	710	0.85	Cast, cold-roll, machine
Vicalloy	52Co, 9.5V	9,600	500	2.80	Cast, hot-swage, cold-draw
Silmaual	87Ag, 9Mn, 4Al	500	5,500[f]	0.06	Machine-punch, cold-roll
Vectolite	$30Fe_2O_3$, $40Fe_3O_4$, $26CO_2O_3$	1,600	900	0.50	Sinter[d]

[a] To be taken as typical only; compositions and properties vary somewhat in practice.
[b] Balance iron plus usual small amounts of impurities.
[c] Hardened at proper stage either by manufacturer or in accordance with manufacturer's instructions.
[d] Heat-treated in a magnetic field; has directional properties.
[e] Has directional properties; should be magnetized in direction of rolling.
[f] Intrinsic coercive force, H_{ci}.

treatment for which is prolonged heating at a temperature of the order of 660 to 700°C. The alloys that are hardened by quenching can be forged and machined, but most of the dispersion-hardening alloys cannot be forged or machined and must be cast to shape and can be cut only by grinding.

The important magnetic characteristics of permanent-magnet alloys are indicated by the part of the hysteresis loop lying between the residual induction and the coercive force and a curve plotted between B and the products of corresponding values of B and H for the same part of the loop. These products are proportional to the magnetic energy stored at the given values of induction. The maximum value of the energy product may be taken as an index of magnetic quality. Typical curves for the three classes of permanent-magnet material are shown in Fig. 17.

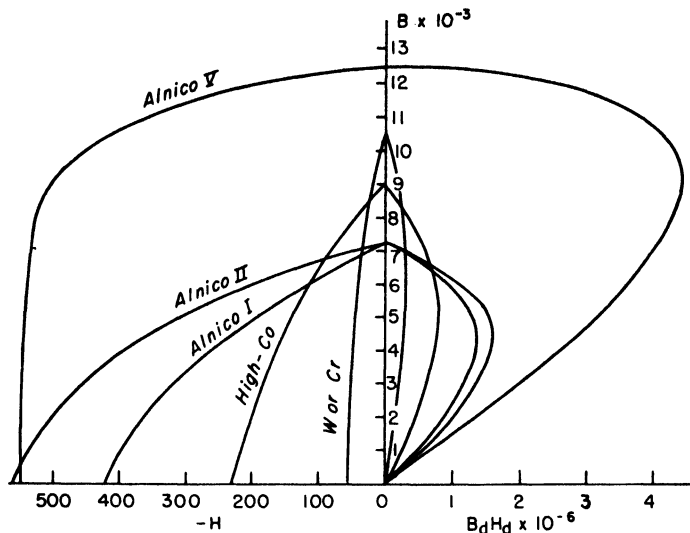



FIG. 17. Demagnetization and energy-product curves for typical magnet materials.

The choice of material for a given application depends upon several factors and, consequently, no one material can be considered as best under all conditions. (See Table 17 for typical characteristics of several permanent-magnet alloys.)

MAGNETIC TESTING

According to the use for which magnetic materials are intended, testing is done by either d-c or a-c methods. In some cases, a material may be subjected to both kinds of test. In the United States, magnetic testing is usually carried out in accordance with the specifications of the American Society for Testing Materials.[1] Since these specifications are modified from time to time to keep pace with the continued improvements in magnetic materials, it is always advisable to consult the latest edition if one contemplates entering the field of magnetic testing. No attempt will be made to describe in detail the techniques involved in magnetic testing as these are to be found in the ASTM specifications. However, there are certain general principles which underlie all techniques and apparatus and which may be of general interest.

54. Tests with Direct Current. The methods and apparatus currently approved as standard by the ASTM for d-c magnetic testing include the Rowland ring method,[2] the Burrows compensated double yoke,[3] the MH permeameter,[4] the Fahy Simplex permeameter,[5] the High-H permeameter,[6] the Babbitt permeameter,[7] the Simplex super-H adapter,[8] and the Saturation permeameter.[9] Some producers have adopted certain nonstandard methods which are simpler and easier to operate and are sufficiently consistent to be used for quality control.

[1] "Ferrous Metals," latest ed., ASTM Standards.

[2] Ewing, J. A., "Magnetic Induction in Iron and Other Metals," 3d ed., p. 138, Electrician Printing Pub. Co., Ltd., London, 1900.

[3] Burrows, C. W., The Determination of the Magnetic Induction in Straight Bars, *Nat. Bur. Standards Bulletin*, S117, **6**, 31, 1909.

[4] Sanford, R. L., and P. H. Winter, A Permeameter for Magnetic Testing at Magnetizing Forces up to 300 Oersteds, *J. Research Nat. Bur. Standards*, **45**, 17, 1950.

[5] Fahy, F. P., A Permeameter for General Magnetic Analysis, *Chem. & Met. Eng.*, **19**, 339, 1918.

[6] Sanford, R. L., and E. G. Bennett, An Apparatus for Magnetic Testing at Magnetizing Forces up to 5,000 Oersteds, *J. Research Nat. Bur. Standards*, **23**, 415, 1939.

[7] Babbitt, B. J., An Improved Permeameter for Testing-Magnet Steel, *J. Optical Soc. Am.* **17**, 47, 1928.

[8] Sanford, R. L., Magnetic Testing, *Nat. Bur. Standards Circ.*, C456, 1946.

[9] Gokhale, S. L., The Saturation Permeameter, *J. AIEE*, **47**, 196, 1928.

The d-c methods listed above are called ballistic methods because they employ ballistic galvanometers (or fluxmeters) in the measurements. Figure 18 is the basic diagram of connections for ballistic tests. The differences between different methods lie in the type of magnetic circuit employed and the arrangement of the magnetizing and test coils. In some types of apparatus, for example, the coil L_2 is located near the surface of the specimen (designated as T in the diagram) and is used to measure the magnetizing force. In other methods such as the Rowland ring or the Burrows compensated double yoke this coil is absent and the magnetizing force is measured in terms of the current in the primary winding. *BG* represents the ballistic galvanometer. The diagram shows alternative ways of connecting the galvanometer. In the one at the right, R_p is the resistance required for critical damping. Some operators prefer this connection. The mutual inductor M_1M_2 is used to calibrate the galvanometer. An excellent treatment of the theory and use of the ballistic galvanometer is given by Harris.[1]

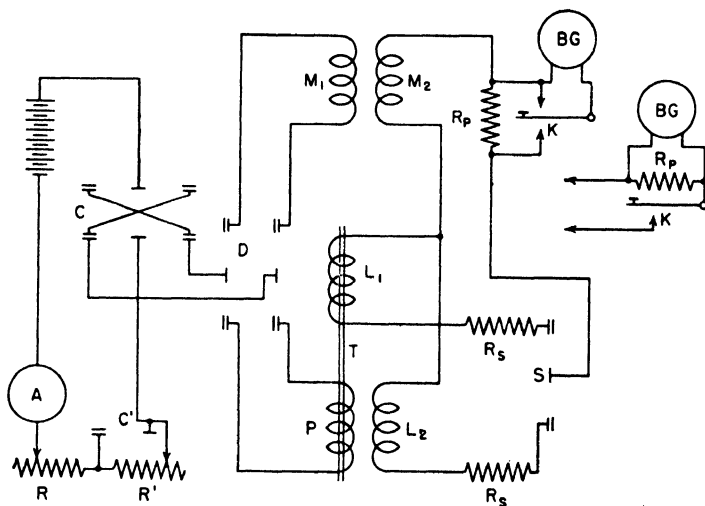



FIG. 18. Connections for ballistic test.

55. Tests with Alternating Current. Several a-c tests for the determination of core loss, a-c permeability, and incremental permeability are specified by the ASTM but some a-c tests required for materials used at higher-than-power frequencies have not at this writing been standardized by the society. However, work is in progress and standard methods should become available in the near future.

Figure 19 shows the basic circuit for the measurement of core loss and a-c permeability on flat-rolled materials. For several years the standard sample was made up of strips 3 cm wide and 50 cm long. The sample weighed 10 kg. These strips were assembled in four equal bundles and inserted in the four coils of an "Epstein" test frame with butt joints at the corners. The coils consisted of two windings, a primary winding and a secondary winding having equal numbers of turns uniformly distributed on the coil forms. In later years, the 50-cm sample has been replaced for the most part by the so-called 25-cm Epstein test. The samples for this test originally consisted of strips 25 cm long and arranged in the test frame to have alternate butt and lap joints at the corners. The present standard specifies double-lap joints and consequently a length of at least 28 cm is required. For details of procedure, for this test, the latest specifications of the ASTM should be consulted.

[1] Harris, F. K., "Electrical Measurements," Wiley, 1952.

One of the permeameters which has not yet been standardized by the ASTM is the radio-frequency permeameter described by Haas.[1] It is a modification of a permeameter originally described by Kelsall[2] for a-c measurements at small magnetizing forces. The instrument was designed primarily for testing toroidal cores without the necessity of placing a winding on each individual core. By an ingenious construction, it is possible to insert a toroid in the instrument and then to surround the core by a single-turn coil by means of a cover plate. The specimen with its single-turn winding then constitutes the secondary of a transformer. The measurements are made in terms of the influence of the sample on the measured impedance of the primary winding of the transformer. The measuring instrument is usually an impedance bridge or a Q meter. Measurements of initial permeability and dissipation factor can be made by this instrument at radio frequencies up to 20 Mc.

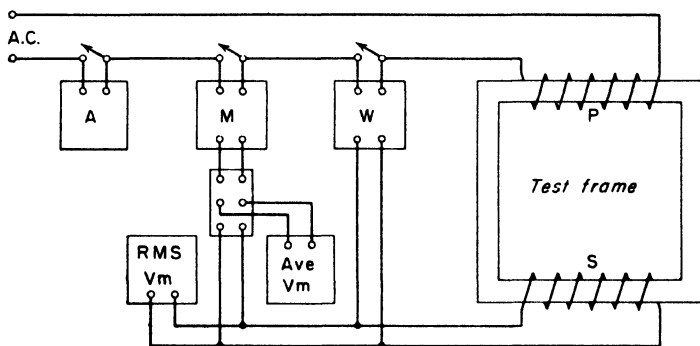



FIG. 19. Circuit for core-loss and a-c permeability tests.

For determining the "dynamic" hysteresis loops, several circuits involving the use of cathode-ray oscilloscopes have been developed. Some of these have been described by Wiegand and Hansen,[3] Zemsky,[4] and others.

56. Tests of Permanent Magnets. The testing of permanent-magnet materials for determining their demagnetization curves is pretty well standardized[5] and need not be discussed here. However, there are some principles involved in the testing of finished magnets that merit some consideration.

It is a quite common practice to judge the strength of a magnet roughly by the force of attraction between it and a piece of soft iron. If the purpose for which the magnet is to be used is to attract or hold an armature, this is a logical method of test, and the attractive force may be measured quantitatively for the purpose of comparing magnets of the same size and shape. However, if the magnet is to be used for any other purpose, the application of such a test may not only give unsatisfactory results but also may prove to be detrimental to the magnet. The characterization of a permanent-magnet material by stating the weight a magnet can lift as related to its own weight is usually misleading, because this relationship is a function not only of the magnetic quality of the material but also of the size and shape of the magnet and the nature of the contact between the magnet and the mass lifted.

Magnets are customarily tested by either one of two methods. One method consists in determining the flux in the neutral zone by quickly removing a test coil con-

[1] Haas, P. H., A Radio-frequency Permeameter, *J. Research Nat. Bur. Standards*, **51,** 221, 1953 (R.P. 2454).

[2] Kelsall, G. A., Permeameter for Alternating Current Measurements at Small Magnetizing Forces, *J. Optical Soc. Am.* and *Rev. Sci. Instruments*, **8,** 334, 1924.

[3] Wiegand, D. E., and W. W. Hansen, A 60-cycle Hysteresis Loop Tracer for Small Samples of Low-permeability Material, *Trans. AIEE*, **66,** 1947.

[4] Zemsky, J., Quantitative Determination of Magnetic Properties by Use of Cathode-ray Oscilloscope, *AIEE Tech. Paper* 47-119, April, 1947.

[5] Sanford, R. L., Magnetic Testing, *Nat. Bur. Standards Circ.* C456, 1946.

nected to a ballistic galvanometer or fluxmeter. The other method consists in determining the field strength in the gap of special pole pieces to which the magnet is applied. The field strength may be measured in terms of the deflection of a pivoted coil in which a definite current is flowing, the emf induced in an armature rotated at a definite speed, or by a test coil and ballistic galvanometer. A bismuth spiral might also be used, but this is less satisfactory than any of the other methods.

Unless the special pole pieces have the same permeance as the pole pieces to which the magnet is to be applied in use, neither of these methods duplicates the conditions under which the magnet is to function and, consequently, the results must be interpreted with care. If the magnet is designed in such a way as to require magnetization with pole pieces in place, with subsequent magnetic stabilization by partial demagnetization, it will be necessary to remagnetize after testing. Otherwise, the required strength will not be obtained in the finished apparatus. Furthermore, a single test will give no indication of whether or not the magnet has been properly aged. Thermal stabilization must be applied to the finished apparatus after all other treatments have been applied, and tests as to its effectiveness must be made by temperature tests on the finished apparatus.

57. Applications. The principal applications involving the use of magnetic materials in radio or other electronic devices are (1) electromagnets, (2) transformers and inductors, (3) permanent magnets.

1. The principal application of electromagnets is in relays.[1] Unless a relay is to be operated with a.c., the core material is usually of solid material. If a.c. is used, the core must be laminated to minimize the demagnetizing effect of induced eddy currents. There are many types of relay in use, and the design required depends upon the functions and conditions of use. If the relay must operate with a minimum of power and can operate at a moderate flux density, some of the nickel-iron alloys may be best to use for core and armature material. However, if higher flux densities can be employed to advantage, a low-silicon iron alloy known as "relay iron" is often used. The silicon tends to decrease aging and the tendency to "stick" after prolonged operation.

Loud-speaker magnets constitute another application of electromagnets. The voice coil of a loud-speaker operates in the air gap of an electromagnet energized by a constant d.c. Solid core materials are used in such electromagnets. In both relays and loud-speaker magnets, the aim of the designer is to keep as much of the total magnetic flux as possible in the air gap where the work is done and so design the magnetic structure that there will be a minimum of stray or leakage flux in other parts of the magnetic circuit. Since there is no insulator of magnetic flux, this involves proper shaping of the parts of the magnetic circuit and location of the magnetizing winding.

Another type of electromagnet that is coming into use is the recording, playback, and erasing heads used in magnetic recording. The magnetic core of such a head is usually constructed of one of the high-permeability nickel-iron alloys. The most common forms utilize the fringing flux across the very short air gap (of the order of 0.001 in.) for recording and erasing. A high value of initial permeability is particularly important for playback heads.

2. The core material for power transformers[2] is usually one of the various grades of silicon steel, the particular grade selected depending on the type of service, whether continuous or intermittent. If a minimum of distortion is required as in interstage coupling transformers, it is necessary either to make the core of relatively large cross section if silicon steel is used or to employ one of the high-permeability nickel-iron alloys. Considerations of space and cost usually govern the choice.

The inductance of coils having magnetic cores is not constant but depends upon the current. The variation is minimized either by an air gap in the magnetic circuit or by the use of powdered materials. In either case the effective permeability of the core material is considerably reduced, but the inductance is still much greater than could be obtained with the same winding without the magnetic core.

[1] Peek, R. L., Jr., and H. N. Wagar, "Switching Relay Design," Van Nostrand, 1955. Packard, C. A., "Relay Engineering," Struthers-Dunn, Inc., 1945. Fisher, R. T., Graphic Solution of Design Problems Involving Sensitive Relays, *Electronics*, October, 1943. Schoemann, A. E., Selecting the Right Relay, *Elec. Mfg.*, June, 1936.

[2] Massachusetts Institute of Technology, "Magnetic Circuits and Transformers," Wiley, 1943.

3. Permanent magnets[1] are used for a great variety of purposes, chiefly in indicating meters, loud-speakers, and polarized relays. The choice of material for a given purpose depends upon a number of factors such as space requirements, degree of stability required, and cost. In many applications where space and weight are not important, the older types of tungsten or chromium magnet steels are satisfactory and more economical in cost than the newer alloys. In applications where space requirements are important and cost is a secondary consideration, the newer alloys such as the Alnicos, Cunife, Cunico, or Vectolite are usually employed. Sometimes the more expensive alloys will have to be used because no other materials are capable of giving the required performance.

[1] Sanford, R. L., Permanent Magnets, *Nat. Bur. Standards Circ.* C448, 1944. Underhill, E. M., Permanent Magnet Design, *Electronics*, December, 1943, January, February, April, 1944.

CHAPTER 4

CAPACITANCE

By E. L. Hall[1] and Philip J. Franklin[1]

1. Capacitance. Capacitance is one of the three electrical quantities present in all radio circuits. The radio engineer endeavors to concentrate capacitance in definite well-known forms at definite points in the circuits, but capacitance exists between conductors in the circuits at different potentials and between the various conductors and the ground. Such capacitances, usually small, are ordinarily of no importance in the case of l-f or a-f currents but may be of great consequence in r-f circuits.

A *capacitor* is an electrical device in which capacitance plays the main role. Although some inductance and some resistance are present, these quantities are often of minor importance and sometimes may be neglected.

A capacitor has three essential parts, two of which are usually metal plates separated or insulated by the third part, called the *dielectric*.

The amount of electricity that the capacitor will hold depends on the voltage applied to it. This may be expressed as $Q = C \times V$. The *capacitance* is the ratio of the quantity of electricity and the potential difference or voltage, or $C = Q/V$, where Q is given in coulombs, C in farads, and V in volts. The capacitance is dependent on the size and spacing of the plates and the kind of dielectric between them.

2. Units of Capacitance. The unit of capacitance is the *farad*. A capacitor has a capacitance of one farad when one coulomb of electricity can be added to it by an applied voltage of one volt. As this unit is too large for practical use, a smaller unit, the microfarad, abbreviated μf, or one-millionth of a farad, is used. Capacitors for radio circuits usually have capacitances between a few thousandths and a few millionths of a microfarad. Another unit, the micromicrofarad, is often used. It is abbreviated $\mu\mu$f. This unit is also called a *picofarad*.

Another unit of capacitance sometimes used is the *centimeter*. The centimeter is equal to 1.1124 $\mu\mu$f.

3. Electrical Energy of Charged Capacitor. Work is done in charging a capacitor because the dielectric opposes the setting up of the electric strain or displacement of the electric field in the dielectric. The energy of the charging source is stored up as electrostatic energy in the dielectric.

The work done in increasing the charge in a capacitor is

$$dW = V\,dQ = \frac{Q}{C}\,dQ \qquad \text{or} \qquad W = \frac{1}{C}\int Q\,dQ = \frac{Q^2}{2C} = \frac{CV^2}{2}$$

where W is expressed in joules, Q in coulombs, and V in volts. The work done in charging the capacitor is independent of the time taken to charge it.

4. Power Required to Charge Capacitor. The average power required to charge a capacitor is given by the equation

$$P = \frac{1}{2}\frac{CV^2}{t}$$

where P is expressed in watts, C in farads, V in volts, and t in seconds. If the capacitor is charged and discharged N times per second, the above equation becomes

$$P = \tfrac{1}{2}CV^2N$$

[1] Diamond Ordnance Fuze Laboratories.

If an alternating emf of frequency f is used in charging the capacitor, the equation may be written

$$P = CE_0^2f$$

where P = power, watts
C = capacitance, farads
E_0 = maximum value of voltage
f = frequency, cps

5. Dielectric Materials. The dielectric of a capacitor may be found in solid, liquid, or gaseous form or in combinations of these forms in a given capacitor.

The simplest form of capacitor consists of two conducting electrodes or plates separated by air. This represents a capacitor having a gaseous dielectric. If this imaginary capacitor has the air between the plates replaced by a nonconducting liquid, such as transformer oil, and if the distance between the plates is the same as in the first case, it will be found that the capacitance has increased several times because the oil has a higher value of *dielectric constant* than air, which is usually taken as 1.

If the space between the plates is occupied by a solid insulator, a capacitor would result, which would be practical as far as the possibility of constructing it is concerned. It would be found, in this case also, that the capacitance of the capacitor was several times larger than when air was the dielectric.

The mechanical construction of either air or liquid dielectric capacitors requires the use of a certain amount of solid dielectric for holding the two sets of plates.

There are a great many dielectric or insulating materials available from which the engineer may choose. It is often found that a material that is very good from the electrical standpoint is poor mechanically, or vice versa. Air is the gas generally used as a dielectric. Compressed air has been used in some high-voltage capacitors, and compressed nitrogen, carbon dioxide, and sulfur hexafluoride are also in use.

Several kinds of oil have been used in capacitors, such as castor oil, cottonseed oil, and transformer oil. Electrolytic capacitors are in wide use in radio equipment as filters and by-pass capacitors where a large capacitance in compact form is required and either a d.c. or pulsating d.c. is applied.

Among the solids used as the dielectric are mica, ceramic materials, and paper. Solid insulators used as mechanical supports include quartz, glass, Isolantite, porcelain, mica, amber, hard rubber, Victron, and Mycalex.

6. Dielectric Properties of Insulating Materials.[1] Such properties as surface and volume resistivity, dielectric strength or puncture voltage, dielectric constant, and absorption are often considered in d-c and commercial-frequency applications. Such data are of little value if the insulating material is to be used at radio frequencies. For the latter application r-f measurements of various properties of the material are essential. A material that may be a satisfactory insulator for low frequencies may be worthless as an insulator at radio frequencies.

One of the most important properties of an insulator for radio frequencies is its power loss. This includes several factors which are difficult to separate but together indicate its suitability for radio purposes. The general idea of the imperfection of a capacitor is brought out in several names such as "power loss," "power factor," and "phase difference," but they are not identical terms.

Dielectric constant is another important property of a material that has a definite bearing upon its use at radio frequencies.

Neither power loss nor dielectric constant alone can be used in selecting the best insulator for a particular application at radio frequencies.

7. Dielectric Constant. The *dielectric constant* K of an insulating material is the ratio of the capacitance C_x of a capacitor using the material as the dielectric to the capacitance C_a using air as the dielectric, or $K = C_x/C_a$. This property of the material is sometimes called *inductivity* or *specific inductive capacity*.

The dielectric constant of a material is not a constant in the true sense of the word,

[1] Von Hippel, A., "Dielectrics and Waves," Wiley, 1954; "Dielectric Materials and Applications," Wiley, 1954.

but varies with the frequency, moisture content, temperature, voltage applied, and manner of applying it. For this reason the term "relative permittivity" is preferred by many engineers.

Table 1 gives the dielectric constants of a large number of electrical insulating materials.

8. Power Loss, Phase Difference, Power Factor, and Loss Factor. Electrical insulating materials are not perfect in their insulating qualities, and a certain amount of power is absorbed in them when used in an a-c circuit. A measurement of the power loss is the best single property that gives an indication of the suitability of an insulating material for use in radio circuits. Power loss can be expressed by a number of quantities, the most commonly used being resistance, power factor, phase difference, and phase angle.

When a.c. flows in a capacitor, the voltage across the capacitor lags somewhat less than 90 deg behind the current as shown by the angle θ (Fig. 1), called the *phase angle*. The complement ψ of the phase angle is called the *phase difference*. The cosine of the phase angle is called the *power factor*. The power loss in the insulating material is

$$P = EI \cos \theta \qquad \text{or} \qquad P = EI \sin \psi$$

where E = voltage across the capacitor
I = current through the capacitor, amp
θ plus ψ = 90 deg, as shown in Fig. 1

From the above, $\sin \psi = \cos \theta$, or the sine of the phase difference is equal to the power factor.

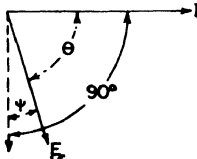



Fig. 1. Phase in a capacitive circuit.

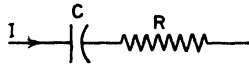



Fig. 2. Capacitance with dielectric losses.

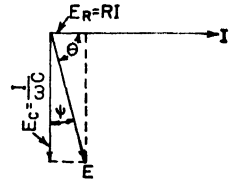



Fig. 3. Vector relations in a capacitor with dielectric losses.

When considering a capacitor having dielectric losses, such as current leakage, brush discharge or corona, dielectric absorption or resistance in the plates, joints, contacts, and leads, it is customary to think of it as a perfect capacitor C with a resistance R in series as shown in Fig. 2.

The voltage vectors may be shown as in Fig. 3, where the resultant voltage E flowing in the circuit is obtained by completing the vector diagram. The angle ψ is quite small for materials suitable for r-f insulators. For small angles the angle $\psi = \tan \psi$. In Fig. 3,

$$\tan \psi = \frac{RI}{I/\omega C} = R\omega C = 2\pi fRC$$

If the resistance, capacitance, and frequency can be measured, the phase difference can be calculated from

$$\psi = 2\pi fRC$$

where ψ = phase difference, radians
f = frequency, cps
R = resistance, ohms
C = capacitance, farads

The following equation is sometimes convenient when wavelength in meters is given:

$$\psi = 0.1079 \frac{RC}{\lambda}$$

where ψ = phase difference, deg
R = resistance, ohms
C = capacitance, $\mu\mu$f
λ = wavelength, m

For small angles, phase difference in radians is equal to power factor (nearly).

Power factor in per cent is 1.745 times phase difference in degrees. *Power factor in* per cent is given by the following equation:

$$\cos\theta = 2\pi fRC \times 10^{-7}$$

where $\cos\theta$ = power factor, per cent
f = frequency, kc
R = resistance, ohms
C = capacitance, $\mu\mu$f

The leakage of electricity by conduction through the dielectric or along its surface contributes to the phase difference but is generally negligible at high frequencies. A capacitor having leakage may be represented by a perfect unit capacitor with a resistance in parallel as shown in Fig. 4. The current divides between the capacitance

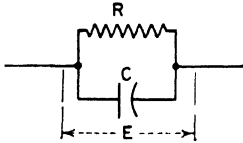



Fig. 4. Equivalent of capacitor with leakage.

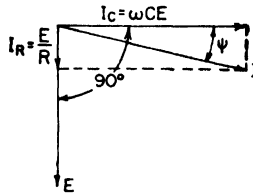



Fig. 5. Vectors in capacitor with leakage.

and the resistance, I_R through the resistance being in phase with the applied voltage E, and I_C through the capacitance leading E by 90 deg as shown in Fig. 5. The resultant current I leads E by (90 deg $- \psi$), where ψ is the phase difference. In Fig. 5,

$$\tan\psi = \frac{E/R}{\omega CE} = \frac{1}{\omega RC} \quad \text{or} \quad \tan\psi = \frac{1}{\omega RC}$$

Power factor is a term that involves all the power losses in a capacitor. If the total power loss in a capacitor is W watts, the voltage applied to it is V volts (rms), and the current flowing through it is I amp (rms), the power factor of the unit is W/VI. The relation between I (amp) and V (volts) for a unit of capacitance C (μf) operating at a frequency f is

$$I = \frac{2\pi fCV}{10^6} = \frac{\omega CV}{10^6}$$

The power factor in per cent may be written

$$\cos\theta = \frac{W \times 10^6}{2\pi fCV^2} = \frac{W \times 10^6}{\omega CV^2}$$

Referring again to Fig. 2, showing the perfect capacitor C and resistance R replacing the actual capacitor, the value of R can be calculated from the equation $W = I^2R$. The quantity R is known as the *equivalent resistance* of the capacitor at the given frequency.

The expression $W \times 10^6/\omega CV^2$ for power factor can be changed into the expression involving resistance, capacitance, and ω by substituting I^2R for W and then substituting $\omega CV/10^6$ for I, giving power factor equal to

$$RC\omega \times 10^{-6}$$

Table 1. Values of Dielectric Constant and Power Factor for Electrical Insulating Materials at Radio Frequencies

Material	Frequency, Mc	Dielectric constant	Power factor	Source
Alsimag No. 243, Forsterite	1	6.2	0.0004	1, 3
	10	6.2	0.0003	1, 3
	100	6.1	0.0003	1, 3
Alsimag No. 196, Steatite	1	5.8	0.0021	1, 3
	10	5.7	0.0015	1, 3
	100	5.6	0.0014	1, 3
Alsimag No. 211, Steatite	1	5.8	0.0004	1
	10	5.7	0.0003	1
	60	4.4	0.0004	2
	100	5.7	0.0003	1
	120	4.4	0.0005	2
Alsimag No. 222, Steatite	1	5.5	0.0002	1
	10	5.5	0.0002	1
	100	5.5	0.0002	1
Amber	0.3	....	0.0051	4
	0.6	....	0.0050	4
	1.0	....	0.0048	4
Aniline formaldehyde resin	1	3.5	0.007	6
	100	3.4	0.004	6
Casein	1	6.2	0.052	6
Celluloid	1	6.2	0.05–0.10	5
Cellulose acetate (plastic)	1	3.9	0.039	6
	100	3.4	0.039	6
Cellulose acetobutyrate	1	3.2	0.017	6
	100	3.0	0.019	6
Dielectene	3,000	3.39	0.0038	7
Epoxy	0.1	3.7	0.02	13
	1.0	3.6	0.02	13
	10.0	3.5	0.02	13
	100.0	3.3	0.03	13
Ethyl cellulose	1	3.4	0.028	6
	100	3.2	0.024	6
Glass:				
Borosilicate No. 707	60	3.7	0.0012	2
	120	3.7	0.0012	2
Borosilicate	18	5.1	0.0059	8
Cobalt	0.5	7.3	0.007	9
Corning No. 707	1	4.0	0.0008	6
	100	4.0	0.0012	6
No. 774	1	5.2	0.0048	6
	100	5.0	0.008	6
No. 790	1	3.9	0.0006	6
	100	3.9	0.0006	6
No. 7052	1	5.1	0.0024	6
	100	5.1	0.0036	6
Nonex No. 772	60	4.2	0.0028	2
	120	4.2	0.0025	2
Plate	0.1	....	0.0077	10
	0.5	....	0.0066	10
	1	....	0.0062	10
Soda lime No. 008	60	6.1	0.0106	2
	120	6.2	0.0103	2
Hard rubber	60	3.1	0.0083	2
	120	3.1	0.0084	2
Low-loss type	60	2.9	0.0057	2
	120	3.0	0.0050	2
Isolantite	0.25–1.5	6.1	0.0018	1
Italian lavite	60	4.7	0.0015	2
	120	4.8	0.0017	2
Lava, grade 1137	1	5.6	0.0040	1
	10	5.5	0.0035	1
	100	5.4	0.0030	1
Lucite HM119, methyl methacrylate	1	2.6	0.015	6
	100	2.6	0.007	6
Lucite	1,333	2.58	0.0090	7
	3,000	2.56	0.0087	7
Lustron	1	2.5–2.6	0.0001–0.0004	1
Melamine formaldehyde resin	1	4.5	0.08	6
	100	4.5	0.03	6
Mica, clear, muscovite	1	6.5–8.0	0.0001–0.0006	11
India, muscovite	0.1–1.0	7.9–7.1	0.0002–0.0001	12
U.S., muscovite	0.1–1.0	8.7–6.6	0.0004–0.0001	12
Mineral oil	1	2.7	0.0008	11
Mycalex, Supramica 560	0.1	7.1	0.002	13
Mycalex	1.0	8.0	0.002	11
Mycalex, red	1,333	5.91	0.0030	7
White	1,333	5.74	0.0033	7
Mycalex 364	1	7.0	0.0021	6
	100	7.0	0.0022	6

Table 1. Values of Dielectric Constant and Power Factor for Electrical Insulating Materials at Radio Frequencies (*Continued*)

Material	Frequency, Mc	Dielectric constant	Power factor	Source
Mykroy 38	1	6.68	0.0016	1
	10	....	0.0015	1
Nylon FM-1	1	3.6	0.020	6
	100	3.6	0.018	6
Paraffin oil	1	2.2	0.0001	6
	100	2.2	0.0004	6
Phenol formaldehyde resins:				
Cast	1	8.0	0.05	6
	100	8.0	0.08	6
General-purpose	1	4.5	0.014	6
	100	4.0	0.014	6
Mineral-filled	1	4.4	0.006	6
	100	4.3	0.012	6
Phenol furfural resins	1	5.0	0.04	6
	100	4.0	0.05	6
Plexiglas, methyl methacrylate	1	2.6	0.015	6
	100	2.6	0.007	6
Polybutene	0.1	2.2	0.006	13
Polyethylene	1	2.25	0.0003	6
	100	2.25	0.0003	6
Polyethylene terephthalate	0.1	3.1	0.01	1
	1.0	3.1	0.01	1
Polyindene	1	3.0	0.0004	5
Polyisobutylene MW 100,000	1	2.22	0.0003	6
	100	2.22	0.0004	6
Polystyrene	60	2.6	0.0005	2
	120	2.5	0.0007	2
Polystyrene MW 80,000	1	2.53	0.0002	6
	100	2.52	0.0003	6
Polyvinyl carbazole	1	2.95	0.0005	6
	100	2.95	0.0006	6
Polyvinyl chlor-acetate	1	2.9	0.014	6
	100	2.8	0.009	6
Polyvinyl chloride	1	2.9	0.016	6
	100	2.9	0.008	6
Polyvinylidine chloride-saran	1	3.0	0.046	6
	100	2.8	0.014	6
Porcelain, wet-process	1	6.5-7.0	0.006-0.008	11
Zircon	1	9.2	0.0010-0.0014	3
Quartz (fused)	1	4.1	0.0002	11
	1	3.8	0.0002	6
	100	3.8	0.0002	6
Clear	60	3.8	0.0003	2
	120	3.8	0.0005	2
Milky	60	3.5	0.0003	2
	120	3.5	0.0005	2
Shellac	1	6.0	0.07	5
	1	3.5	0.031	6
	100	3.1	0.030	6
Styraloy 22	1	2.4	0.0012	6
	100	2.4	0.0043	6
	1,333	2.40	0.0047	7
Styramic	1	2.75	0.0002	6
	100	2.73	0.0002	6
Styramic HT	1	2.64	0.0002	6
	100	2.62	0.0002	6
Styramic E 1689	3,000	2.55	0.00087	7
Tenite II	3,000	2.95	0.031	7
Textolite No. 1422	3,000	2.4-2.5	0.0006-0.0009	1
Urea formaldehyde resins	1	5.6	0.028	6
	100	5.0	0.05	6
Urea resin, wood-flour filler	1	5.7	0.030	5
	10	5.5	0.038	5
	35	5.3	0.042	5
Vinylite V	1,333	2.78	0.0076	7
	3,000	2.61	0.0068	7
Wax, beeswax	1	2.9	0.025	11
Boler	1,333	2.17	0.00019	7
Ceresin	1	2.5-2.6	0.0012-0.0021	11
Halowax	1	3.7	0.0014	6
	10	3.4	0.105	6
Paraffin	1	2.5	0.0097	11
	1	2.25	0.0002	6
	10	2.25	0.0002	6
	1,333	2.17	0.00019	7
Superla	1,333	2.26	0.00019	7
	3,000	2.26	0.00015	7

Sources for Tables

1. Manufacturers' bulletins.
2. Miller, J. M., and B. Salzberg: Measurements of Admittances at Ultra-high Frequencies, *RCA Rev.*, **3**, 480, April, 1939.
3. Thurnauer, H.: Low-loss Ceramic Dielectric, *Tele-Tech.*, **6**, 86, February, 1947.
4. Schott, Erich: Hochfrequenzverluste von Gläsern und einigen anderen Dielektrics, *Jahrb. drahtlosen Tele. Tele.*, **18**, 82, August, 1921.
5. Bloomfield, G. F.: Insulating Materials for the Higher Frequencies, *T & R Bull.*, **14**, 635, May, 1939; *Radio Tech. Digest*, September-October, 1939, p. 23.
6. "Reference Data for Radio Engineers," 4th ed., Federal Telephone and Radio Corporation, 1956.
7. Englund, C. R.: Dielectric Constants and Power Factors at Centimeter Wave-lengths, *Bell System Tech. J.*, **23**, 114, January, 1944.
8. Chaffee, J. G.: The Determination of Dielectric Properties at Very High Frequencies, *Proc. IRE*, **22**, 1009, August, 1934.
9. Hoch, E. T.: Power Losses in Insulating Materials, *Bell System Tech. J.*, **1**, 110, November, 1922.
10. MacLeod, H. J.: The Variation with Frequency of the Power Losses in Dielectrics, *Phys. Rev.*, **21**, 53, 1923.
11. General Electric Company.
12. Lewis, A. B., E. L. Hall, and F. R. Caldwell: Some Electrical Properties of Foreign and Domestic Micas and the Effect of Elevated Temperatures on Micas, *Bur. Standards, J. Research*, **7**, 403, August, 1931.
13. Diamond Ordnance Fuze Laboratories.

The power loss or rate at which heat is generated within a dielectric is proportional to the product of the dielectric constant and the power factor. This product is known as the "loss factor" of the dielectric. The power loss in watts per cubic inch may be calculated from the following:[1]

$$P = 2\pi f G^2 K \tau \times 0.2244 \times 10^{-12}$$

where P = power loss, watts per cu in.
f = frequency, cps
G = voltage gradient in dielectric, rms volts per in.
K = dielectric constant
τ = power factor of dielectric

9. Dielectric Properties. Table 1 gives dielectric constant and power factor at certain frequencies of a large number of electrical insulating materials, as obtained from the sources given at the end of the table. In some cases data from different sources do not agree, but differences in composition, method of making measurements, and condition of samples may account for such disagreements.

10. Dielectric Strength. The *dielectric strength* of an insulating material is the minimum value of electric field intensity required to rupture it. Dielectric strength is usually expressed in kilovolts per centimeter of dielectric thickness. The fall in insulation resistance with rise in temperature is a factor of great importance in connection with the breakdown of a dielectric under the applied voltage. Insulating materials are not strictly homogeneous. The current leakage through an insulating material may perhaps be concentrated in a few small paths through the material, and the energy loss due to the leakage, although small, may be large compared with the area through which it is flowing. The paths of the current flowing through the dielectric become heated with a resulting lowering of the resistance of the path and an increase in the current leakage. The heating of the dielectric may lead to rapid deterioration, particularly if moisture is present, and ultimate breakdown. The length of time of the application of the voltage has a definite bearing upon the breakdown voltage. Most dielectrics will withstand for a very brief period a much higher voltage than they can when the voltage is applied for a longer period.

These effects have dictated two tests for capacitors: a high flash-test voltage of very brief duration, and the application of a much lower voltage for a longer period.

The dielectric strength of a material is usually found to be lower for r-f voltages than for a-f or d-c voltages. The rupturing voltage at radio frequencies depends on the rapidity with which the voltage is raised and is not nearly so definite a phenomenon

[1] Hoch, E. T., Power Losses in Insulating Materials, *Bell System Tech. J.*, **1**, 110, November, 1922.

as l-f puncture voltage. Dielectric strength of solid insulators is difficult to measure because of the complexity of the experimental effects. As the r-f currents flow in the material, heating, corona, flashover, and possible deterioration, blistering, or charring may result with consequent changing of voltage and current as the time of application elapses.

If high r-f voltages are applied to an air capacitor, a corona discharge may be set up which appears as a visible glow around high-potential metal parts, points, and sharp edges and is usually distinctly audible. These corona effects represent a power loss. Hence the construction of air capacitors for high voltages requires the rounding of all edges and corners and the avoiding of sharp points which encourage the formation of corona and flashover.

11. Dielectric Absorption. When a capacitor is connected to a d-c source of emf, the instantaneous charge is followed by the flow of a small and steadily decreasing current into the capacitor. The additional charge is absorbed by the dielectric. Similarly the instantaneous discharge of a capacitor is followed by a continuously decreasing current. The capacitor does not become fully charged immediately, nor does it completely discharge immediately when its terminals are shorted, but several discharges may be secured when the capacitor possesses dielectric absorption. The maximum charge in a capacitor cyclically charged and discharged varies with the frequency of charge.

If a capacitor evidencing dielectric absorption is used at radio frequencies, a power loss occurs which appears as heat. The existence of power loss indicates a component of emf in phase with the current as though a resistance were in series with the capacitor as shown in Fig. 2. The effect of dielectric absorption can be measured along with other losses in the capacitor, although dielectric absorption represents the chief power loss in solid dielectrics.

12. Calculation of Capacitance. Formulas are available for use in calculating the capacitance for a large number of geometrical shapes of conducting surfaces such as spheres and cylinders, either separated or concentric, and flat surfaces of various shapes. The usual types of capacitor calculations are concerned with two or more flat conductors.

When two conducting plates are parallel, close together, and of large area, the capacitance is given by

$$C = 0.0885 \times \frac{KS}{t}$$

where C = capacitance, $\mu\mu$f
K = dielectric constant (which is 1 for air)
S = area of one plate, sq cm
t = distance between plates, cm

When more than two plates are used, the formula becomes

$$C = 0.0885 \times \frac{KS(N-1)}{t}$$

where N = number of plates

The actual capacitance of a parallel plate capacitor is slightly larger than the value as calculated from the above formula, because of the fringing of the electric lines of force beyond the space between the plates. A correction[1] can be made for this fringing by slightly increasing the dimensions of the plates. A narrow strip of width w can be added to the actual plate dimensions. In the case of circular plates $w = 0.4413t$, and for plates with straight edges $w = 0.110t$, where t is the distance between the plates in centimeters.

13. Combinations of Capacitors. Combinations of two or more capacitors in a circuit are often arranged in either series or parallel. Perfect capacitors connected in parallel give a total capacitance equal to the sum of the capacitances of the individual capacitors. Capacitors connected in series give a resulting capacitance which may

[1] Coursey, Philip R., "Electrical Condensers," Sir Isaac Pitman & Sons, Ltd., London.

be calculated from the following:

$$C = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots}$$

This formula gives the following expression in the case of two capacitors in series:

$$C = \frac{C_1 \times C_2}{C_1 + C_2}$$

The various elements such as tubes, sockets, mountings, and wiring in radio apparatus contain many small capacitances by virtue of the difference of potential existing between the numerous conductors insulated from one another. These small capacitances are known as *stray* capacitances. Although they are unimportant in some kinds of work, in other types of work, such as in amplifier design, they must be taken into account. In the case of resistance-coupled amplifiers, for example, these capacitances reduce the amplification at the higher audio frequencies and make a flat characteristic with high over-all gain impossible.

The effect of stray capacitances is eliminated in the case of units used as capacitance standards by shielding the insulated plates and grounding the shield. In this manner a definite capacitance is always assured for a given scale setting.

14. Effect of Frequency on Capacitance. In the best capacitors the effect of frequency is nil in the lower radio range. A criterion of a suitable capacitor for use as a standard is that its capacitance shall be the same for two different sets of charging and discharging conditions. A variable air capacitor, such as the National Bureau of Standards type described on p. 120 of the Bureau's *Circ.* 74, gives the same capacitance at 100 and at 1,000 charges and discharges per second. A capacitor having considerable solid dielectric in its make-up will show a difference in capacitance with frequency. The quantity of electricity that flows into a capacitor during a finite charging period is greater than would flow in during an infinitely short charging period. Consequently the measured or apparent capacitance with a.c. of any finite frequency is greater than the capacitance on infinite frequency, the latter being called the *geometric* capacitance. The capacitance decreases as the frequency increases.

FIG. 6. Resonant frequency vs. capacitance for tab paper, metallized, mica, and ceramic capacitors. Deviations from the curve of all types are about the same. (*From Geiser.*)

The length of the internal leads should be kept as short and direct as possible to minimize the inductance of the leads that acts to give an apparent change of capacitance with frequency. The amount of this change can be calculated from $C_a = C(1 + \omega^2 CL \times 10^{-12})$, where C_a is the apparent or measured capacitance, C is in microfarads, and L is in microhenrys.

Resonance is defined as the frequency at which the capacitance and series inductance of a capacitor produce zero reactance. It follows that the resonant frequency can be raised by minimizing lead inductance.

The more nearly opposite the lead connections, the lower the effective series inductance becomes. This is demonstrated by an increase in the two-terminal resonant frequency.

The resonant frequency increases with decreasing capacitance value. A plot of capacitance vs. frequency on loglog paper, shown in Fig. 6, yields a straight line for similar capacitors.

Very little difference is noted[1] in resonant frequency between inserted tab paper, mica, metallized, and ceramic capacitors.

15. Capacitor Types. There are many ways in which capacitors might be classified: by their construction, size, voltage rating, use, dielectric, or fixed or variable capacitance. The capacitors used in various radio applications are found in innumerable sizes, shapes, and uses. *The two simplest divisions into which they may be classified have to do with their capacitance, i.e., whether it is fixed or variable.*

16. Types of Fixed Capacitors. Fixed capacitors are available in all ranges from a few micromicrofarads to several microfarads, for any voltage rating, and in innumerable shapes and dimensions, all depending on the use for which the unit is intended.

The three general methods used in construction of fixed capacitors are: (1) winding, which is used for flexible dielectrics such as impregnated paper, plastic films, and some glasses; (2) stacking, which is used for stiff, brittle dielectrics such as mica, glass, and some combinations of organic and inorganic materials; and (3) molding or stamping, which are used primarily with ceramics which are molded wet into a circular or rectangular plate or tube and then fired.

Wound capacitors are formed by winding two metal-foil electrodes or ribbons in conjunction with one or more dielectric ribbons. The number of layers of dielectric is determined by the desired voltage rating. There are two types of windings, *inserted-tab* and *extended-foil*. In winding the inserted-tab type, the foil used is narrower than the dielectric and contact is made with the foils by inserting one or more tabs per foil during the winding. In the extended-foil type the foils have the same width as the dielectric. The foils are staggered so that they extend over the edges of the dielectric. The terminals are soldered to the extending foil at the opposite ends. The latter type of construction produces capacitors having minimum series resistance. A third type of wound capacitor, employing metalized dielectric, is in common use. In this type a thin aluminum or zinc film from 25 to 100 $m\mu$ in thickness is deposited on the dielectric. This construction increases the series resistance, but for many applications the physical sizes are only about one-third of those of the more conventional designs.

The stack-type capacitor is normally built up by alternating sheets of metal foil and dielectric. The sets of foils of opposite polarity are brought out at opposite ends and soldered together. The whole stack of plates is rigidly clamped in such a way as to grip the plates firmly in the center and expel all air from the stack. A casing of some moisture-resisting insulation, usually phenolic or ceramic, is then molded around the unit.

17. High-voltage Capacitors. It is customary to mount large high-voltage capacitors in steel tanks which are filled with a high-flash-point oil which protects against dirt and moisture, prevents flashover along the capacitor sections, insulates the capacitor from the tank, and conducts heat away from the capacitor elements.

Fixed high-voltage vacuum- or gas-filled capacitors are available for transmitters or power amplifiers. They are normally two cylindrical metal elements mounted in an exhausted glass tube. The gas-filled capacitors usually employ sulfur hexafluoride, argon, or helium. See also Sec. 27.

Fixed vacuum-dielectric capacitors of about 5 to 50 $\mu\mu f$ are used as all or part of plate tank, high-voltage coupling, by-pass, and neutralizing capacitance at high frequency.

18. Ceramic Capacitors. Development of ceramics having high and controllable dielectric constants has been in progress for a number of years and capacitors utilizing the properties of these ceramics are in very wide production and use. Barium-strontium titanates have been found to possess dielectric constants from 10,000 to 18,000. Barium-calcium titanates have lower dielectric constants, and barium-magnesium titanates still lower values. The body is prepared as a wet mix which is subsequently molded or stamped to the desired shape and processed at high tempera-

[1] Geiser, D. T., "An Investigation of Lowest Resonance Frequency in Commercially-available By-pass Capacitors," Convention Record of the IRE, 1954 National Convention, Part 3, p. 43.

tures such as 1250 to 1400°C. Capacitors formed of tiny disks of these materials have been used in the construction of "matchbox" receivers and "lipstick" transmitters in which all components are very small.

The ceramic and silver-mica capacitors have plated-on or fired-on silver electrodes to which the leads are soldered. These are normally protected by either phenolic, epoxy, or ceramic cases.

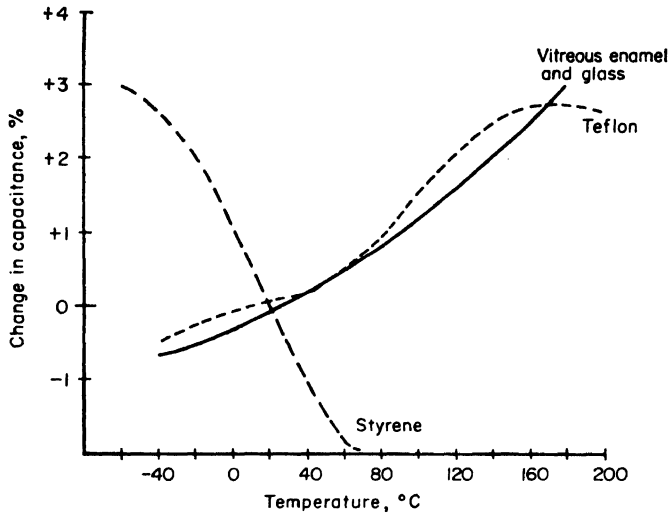



FIG. 7. Capacitance-temperature characteristics of several capacitors.

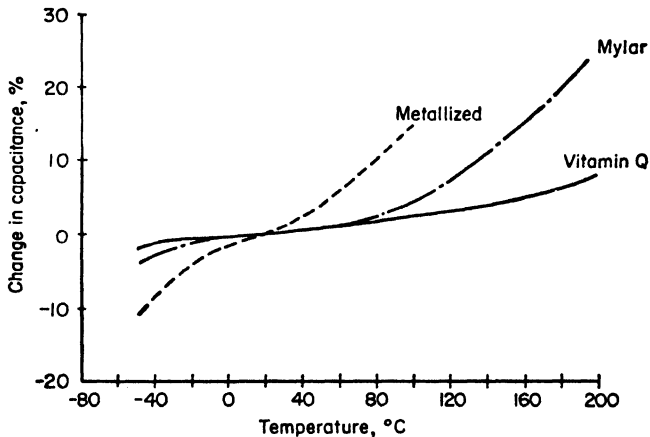



FIG. 8. Dependence of capacitance on temperature for several capacitors

The change of capacitance of a capacitor with temperature is dependent on the dielectric and impregnant used. Figures 7 and 8 show the per cent change in the capacitance with temperature of several capacitors using different dielectrics. In most applications predictable temperature coefficients are not required, provided the total change in capacitance is less than some desired limit. It is for this reason that some capacitors are marketed as GMV (guaranteed minimum value). In some applications, however, such as for tuning capacitors, it is important that the capaci-

tance remain fixed with changes in temperature or change predictably within narrow limits so that they may be used to compensate for warm-up drift in inductors. Ceramic tube or disk capacitors are available for this purpose and are manufactured with predictable positive, zero, or negative temperature coefficients from +100 to −5,600 ppm per °C in a capacitance range from 0.3 to 4,500 μμf.

Where high stability with respect to temperature is required, silver-mica capacitors have proved to be very useful. They suffer somewhat from the tendency for the silver coating to migrate in the presence of moisture and with d-c voltages impressed across the terminals.

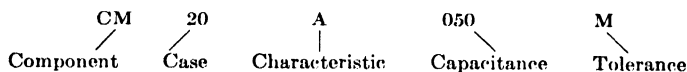
19. Type Designations.[1] RETMA standards and MIL specifications make use of type designations to describe capacitors and their properties fully in the form of a code containing numerals and letters. An example is given below:

Component Designation. Fixed mica capacitors are identified by the symbol CM. Fixed paper capacitors are identified by the symbol CP, etc. In RETMA standards the two-letter symbol is preceded by the letter R to indicate RETMA standards are being used.

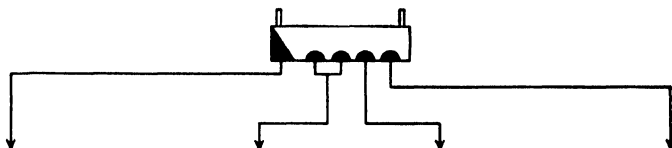
Table 2. 1957 RETMA Standard Color Coding for Molded Mica Capacitors

White for mica

Designation	Q	Insulation resistance not less than	T.C. not more than	Capacitance drift not more than	Color	Tolerance	First and second significant figure of capacitance	Decimal multiplier of capacitance
Class A	a	3,000 megohms	± 1,000 ppm/°C	± (5% + 1 μμf)	Black	20%(M)	0	1
Class B	b	6,000 megohms	+ 500 ppm/°C	+ (3% + 1 μμf)	Brown	1%(F)	1	10
Class C	b	6,000 megohms	± 200 ppm/°C	± (0.5% + 0.5 μμf)	Red	2%(G)	2	100
Class D	b	6,000 megohms	± 100 ppm/°C	± (0.3% + 0.1 μμf)	Orange	3%(H)	3	1,000
Class E	b	6,000 megohms	+100 to −20	± (0.1% + 0.1 μμf)	Yellow	....	4	10,000
					Green	5%(J)	5	
					Blue	....	6	
					Violet	....	7	
					Gray	....	8	
					White	....	9	
					Gold	....	...	0.1
					Silver	10%(K)	...	0.01

a. Q must be greater than 30% of a minimum allowable Q for other characteristics.
b. Minimum acceptable Q between 0.5 and 1.5 Mc is defined by a curve; value varies with capacitance.

Case Designation. The case designation is a two-digit symbol that identifies a particular case size and shape.

Characteristic. The characteristic is indicated by a single letter which is the class of electrical properties into which the unit falls. Reference is made to Table 2 for details.

[1] A considerable amount of information on the numerous types of capacitors now available for commercial and military use will be found in "Electronic Components Handbook," McGraw-Hill, 1957.

Capacitance Value. The nominal capacitance value in micromicrofarads is indicated by a three-digit number. The first two digits are the first two digits of the capacitance value in micromicrofarads. The final digit signifies the number of zeros that follow the first two digits. If more than two significant figures are required, additional digits may be used, the last digit always indicating the number of zeros.

Capacitance Tolerance. The symmetrical capacitance tolerance in per cent is designated by a letter as shown in Table 2.

Color Coding. On most capacitors there is sufficient space available on the body of the unit for all appropriate information to be printed on it. On the small capacitors, however, like mica and ceramic units, a color-coding scheme has been adopted for marking the

Table 3. 1957 RETMA Standard Five-dot Color Coding for Ceramic Dielectric Capacitors

Temperature coefficient of capacitance	Color	First and second significant figure of capacitance	Decimal multiplier of capacitance	Tolerances of capacitance	
				Nominal $\leqq$ 10 μμf	Nominal > 10 μμf
0 ppm/°C	Black	0	1	±2.0 μμf	±20%
−33 ppm/°C	Brown	1	10	±0.1 μμf	± 1%
−75 ppm/°C	Red	2	100	. . . .	± 2%
−150 ppm/°C	Orange	3	1,000	. . . .	± 3%
−220 ppm/°C	Yellow	4	10,000		
−330 ppm/°C	Green	5	. . . . . .	±0.5 μμf	± 5%
−470 ppm/°C	Blue	6			
−750 ppm/°C	Violet	7			
a	Gray	8	0.01	±0.25 μμf	
b	White	9	0.1	±1.0 μμf	±10%

a. This is a general-purpose capacitor having any nominal temperature coefficient between +150 and −1,500 ppm/°C.

b. This is a general-purpose capacitor having any nominal temperature coefficient between 100 and −750 ppm/°C.

required information on each unit. Table 2 gives the 1957 RETMA standard color coding for mica capacitors and Table 3 gives the same information for ceramic capacitors. In the case of tubular ceramic capacitors an alternate coding using bands of colors (Fig. 9) rather than dots is extensively employed. The only requirement is that the first band be distinguishable from the last band. This is usually accomplished by making the first band wider.

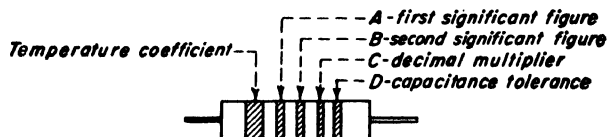



Fig. 9. Capacitor color-code arrangement.

20. Electrolytic Capacitors. Another type of fixed capacitor with important applications is known as the electrolytic capacitor. Its advantages are low cost and high capacitance as compared with other types of fixed capacitors.

The electrolytic capacitor consists of three essential components: the anode, the dielectric film, and the electrolyte. The anode is made of either aluminum or tantalum of high purity and forms one plate of the capacitor. The dielectric film is

formed electrochemically on the anode and is very thin. The electrolyte may be either a liquid, paste, or more recently a solid substance. It is the second plate of *the capacitor, insulated from the anode plate by virtue of the dielectric film formed* on the latter.

According to the electrolyte employed, electrolytic capacitors may be divided into three general classes: (1) dry electrolytic capacitors in which a pastelike form of electrolyte is used, (2) wet electrolytic capacitors in which a liquid or waterlike electrolyte is used, and (3) solid electrolytic capacitors in which a solid electrolyte is used. With the recent development of the solid electrolytic capacitor the designation dry electrolytic capacitor for those units employing paste electrolyte seems improper, but for the purpose of this book this term will be retained.

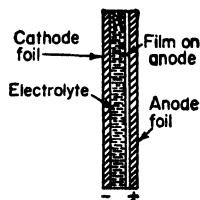



FIG. 10. Electrolytic-capacitor construction.

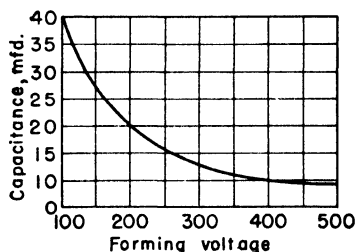
The electrolyte in the case of dry electrolytic capacitors is absorbed in porous paper and held in position adjacent to the anode foil by this paper. In addition another metal foil, generally though perhaps erroneously called the cathode foil, is incorporated for the purpose of making electrical contact to the electrolyte-saturated paper.

In the wet type the electrolyte is a dilute aqueous solution. The anode member with adhering dielectric film is suspended in a can. For aluminum electrolytic capacitors the can is generally aluminum; for tantalum electrolytic capacitors the can is a silver-plated metal such as copper or brass. The can is filled with the electrolyte and acts as the electrical connection to the electrolyte similar to the cathode foil in the case of the dry electrolytic capacitor.

All aluminum electrolytic capacitors have anodes of aluminum foil. Tantalum capacitors may employ tantalum foil, slugs of sintered porous tantalum, or tantalum wire as anode material.

The solid electrolytic capacitor contains no electrolyte. The most commonly used structure consists of the conventional porous tantalum anode, a layer of semiconducting material such as manganese dioxide, a suitable conductor, and a metallic terminal. Recently, solid electrolytic capacitors that use a tantalum wire anode have become available. Foil-type solid electrolytic capacitors, as well as units employing aluminum foil, are under development. In general, the use of solid electrolytic capacitors is increasing.

FIG. 11. Electrolytic-capacitor characteristic.

For a given area of anode surface the capacitance in microfarads is inversely proportional to the thickness of the dielectric film. The film thickness is proportional to the voltage during electroformation of the film. Therefore, units with very low voltage ratings may be made with very high capacitances. If the anode area is such as to give 8 μf when the working voltage is 500 volts d-c, then the same area at lower working voltages will yield a capacitance as indicated on the curve of Fig. 11. The ordinary ranges are 500 to 6,000 μf in capacitance for voltage ratings of 6 to 60 volts d-c and 2 to 100 μf with voltage ratings of 100 to 150 volts d-c.

21. Electrolytic Capacitor Characteristics. The d-c voltage which an electrolytic capacitor can withstand is governed by the voltage at which the original film is applied. For the more common polar-type capacitors it is necessary that the anode always be connected to the positive side of the voltage source. The capacitor connected in this manner will operate satisfactorily as long as the applied voltage is of correct polarity and does not exceed rated voltage for more than a few seconds at a time. A reversal of potential will cause the unit to draw considerable current even at low voltages. A d-c voltage in excess of rated causes the unit to draw appreciable

leakage current. A nonpolar-type electrolytic capacitor is composed of two units mounted back to back in the same case. The capacitor will operate satisfactorily no matter what the polarity if the voltage rating is not exceeded.

Dry electrolytic capacitors have a definite breakdown voltage at which permanent failure occurs. Momentary surges less than breakdown voltage but higher than operating voltage will ordinarily do no damage.

Electrolytic capacitors have a power factor that is considerably higher than other types of fixed capacitors. This is due in part to the fact that one of the conducting plates is the electrolyte, which has considerably higher resistance than the conventional metallic plates of the other types. In effect this places a resistance in series with the capacitor and hence causes a high power factor of the entire unit. Dry electrolytic capacitors have a power factor of about 6 per cent at 60 to 120 cps. Power factors increase with frequency and for this reason electrolytic capacitors are generally confined to l-f applications. Tantalum electrolytic capacitors are superior to aluminum electrolytic capacitors in power factor because the dielectric film in the tantalum units is chemically resistant to a wider range of electrolytes than the film in the aluminum units and lower electrical resistance electrolytes can be employed.

22. Etched-foil Types. Methods of treating the smooth foil surface in such a way as to make it extremely rough have been applied to electrolytic capacitors. The anode then has an increased total area over and above the original smooth surface. The dielectric film follows the contours of the foil, and the result is a great increase in capacitance with no increase in volume for any given working voltage.

Several methods have been evolved for formation of a roughened surface for the anode foil. Chief among these are as follows: (1) etching, whereby the smooth surface is attacked either chemically or electrochemically, and (2) spraying, whereby the metal is sprayed in a molten state on a suitable carrier medium.

23. Effect of Temperature on Characteristics. Electrolytic capacitors operate best under normal conditions of temperature. The limitations of the electrolyte and the film are the governing factors in the operation of this type of capacitor.

Extremely high temperatures cause the electrolyte to dry out and increase in resistivity with consequent increase in power factor when normal temperature is again reached. Furthermore the increase in d-c leakage current with temperature must be considered since there is danger of the start of a destructive cycle due to the generation of internal heat because of the increased d-c leakage.

Low temperature causes a decrease in capacitance and an increase in power factor. This is more pronounced in aluminum than in tantalum capacitors using liquid electrolyte and much more pronounced in both of these than in solid tantalum capacitors. These changes are temporary and are restored to normal when normal operating temperatures are again reached.

When high operating temperatures are to be experienced, the units should be hermetically sealed in metal cans. This construction limits the loss of electrolyte to a minimum, and longer life can be expected.

24. Comparison of Aluminum and Tantalum Electrolytic Capacitors. Tantalum capacitors are superior to aluminum capacitors in most respects. The exceptions are that at present tantalum electrolytic capacitors cannot be formed at so high a forming voltage or operated at so high a working voltage as aluminum capacitors. Single-cell tantalum capacitors employing liquid electrolytes are rarely found rated at higher than 150 volts d-c and the solid tantalum units are yet to be fabricated for ratings above 35 volts d-c. Among the three types of tantalum capacitors containing a liquid electrolyte, the type employing an anode of sintered porous tantalum will operate over the widest temperature range (−55 to 200°C) but this type has the least convenient shape.

The use of etched tantalum foil as anode material allows construction of capacitors having a smaller diameter and greater mechanical stability than porous-plug-type units, but foil-type capacitors cannot be operated at temperatures above 125°C. The use of a wire-type anode allows further savings in space, but the capacitance and voltage ratings of these capacitors are lower at the present time, and their maximum

usable temperature now is 85°C. The solid-electrolyte units available to date cannot be used above 85°C, but their low-temperature characteristics are superior to any electrolytic capacitors containing a liquid electrolyte. Because of the lesser increase in power factor at low temperature, solid-electrolyte capacitors can be used satisfactorily at −80°C, while no liquid-electrolyte unit has been rated for less than

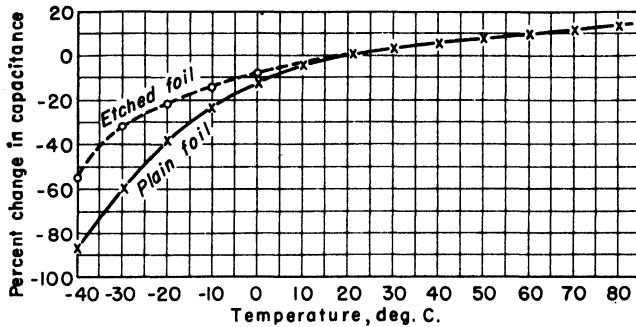



FIG. 12. Temperature characteristics, plain and etched foil.

−55°C. The electrical characteristics of solid-electrolyte capacitors are also less frequency-dependent, so that tests on these units have been extended up to 10 Mc, compared with a maximum of 100 kc for the liquid-electrolyte types. In addition, the storage ability of solid-electrolyte capacitors is greatly superior to that of the other types.

Table 4 lists voltage rating, capacitance, and temperature ratings as well as the dimensions of some characteristic units of each of the described types of electrolytic capacitors. The cost of tantalum electrolytic capacitors is several times that of the aluminum electrolytic capacitors. It is worth noting that tantalum electrolytic capacitors are finding their main application in miniature and subminiature circuits where their high capacity-to-volume ratio and long life meet requirements very nicely.

Table 4. Comparative Size and Performance Data for Five Types of Electrolytic Capacitors*

Type	Voltage rating, volts	Capacitance, μf	Operable temp range, °C	Physical size	
				Length, in.	Diam, in.
Dry Al	450	80	−40 to 85	4.250	1.375
	600	10	−20 to 85	2.938	1.063
	450	10	−40 to 85	1.688	0.875
Porous Ta slug	90	50	−55 to 85	0.688	0.875
	60	50	−55 to 200	0.688	0.750
	90	10	−55 to 85	0.688	0.468
	70	10	−55 to 125	0.688	0.468
Ta foil	150	1.5	−55 to 125	1.031	0.281
	25	10	−55 to 125	1.031	0.281
Ta wire	8	20	−55 to 85	0.500	0.125
	20	1	−55 to 85	0.172	0.095
Solid Ta	20	15	−80 to 85	0.438	0.175
	8	10	−80 to 85	0.312	0.125
	35	1	−80 to 85	0.250	0.125

* Compiled from several manufacturers' specification sheets.

Literature on Tantalum Capacitors

Foster, Levin W.: Tantalytic Capacitors, *Gen. Elec. Rev.*, **54**, 30, October, 1951.

Güntherschulze, A., and Hans Betz: "Elektrolytkondensatoren," 2d ed., Herbert Cram, Publisher, 1952.

Hall, J. H., and Fred H. Berghorn: Miniaturizing the Tantalum Capacitor, *Elec. Mfg.*, **46**, 82, December, 1950.

Hovey, R. J., and S. S. Fry: Tantalum and Tantalum Capacitors, *Elec. Mfg.*, **57**, 80, March, 1956.

McLean, D. A., and F. S. Power: Tantalum Solid Electrolytic Capacitors, *Proc. IRE*, **44**, 872, July, 1956.

Peck, D. B., S. W. Bubriski, and W. W. Schroeder, Jr.: High-temperature, Foil-type Tantalum Capacitors, *Elec. Mfg.*, **57**, 134, May, 1956.

Warner, D. F.: Application of Tantalum Electrolytic Capacitors, *Trans. IRE Group on Component Parts*, **7**, November, 1955.

Whitehead, M.: Tantalum Electrolytic Capacitors, *Bell Lab. Record*, **28**, 448, October, 1950.

25. Applications. The nature of electrolytic capacitors makes them particularly suitable for filter circuits in power supplies where a relatively high capacitance is required together with the ability to withstand a d-c potential and small superposed a-c ripple. Second only in importance is the use as a-f by-pass capacitors across screen grids and cathode bias resistors. The use of a-c electrolytics wherein the cathode foil is replaced with a second anode is important in capacitor motor service. These latter units are divided into two classes: those for intermittent duty and those for continuous duty. The former are rated at from 30 to 500 μf at 100 volts a.c. and the latter at 10 to 50 μf at 25 volts a.c.

The intermittent-duty type functions only during the starting of a capacitor motor (capacitor start-induction run). The continuous-duty type functions in the smaller motors rated at about 1/100 hp and is continuously on voltage during the operation of the motor.

26. Testing. The circuit of Fig. 13 is generally used to test electrolytics in production. E_{dc} supplies a polarizing voltage so that the voltage across the capacitor will be pulsating d.c. The isolating capacitor prevents short-circuiting the polarizing voltage. If E_{dc} is maintained at a constant value, the a-c milliammeter may be calibrated in terms of the capacitance under test. I_{dc} reads the d-c leakage current.

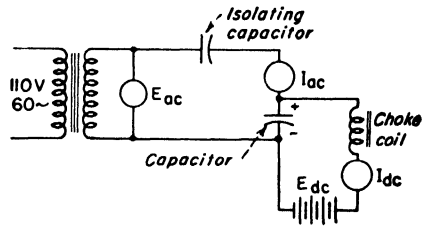



FIG. 13. Production testing circuit for electrolytic capacitors.

For the accurate measurement of capacitance and power factor, bridge systems such as those shown in Fig. 14 or 15 should be used. They are essentially standard bridge systems rearranged to permit the application of a polarizing voltage.

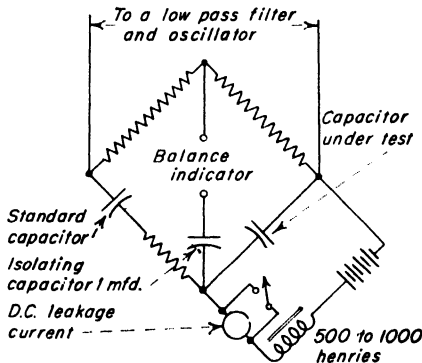



FIG. 14. Capacitor measuring circuit.

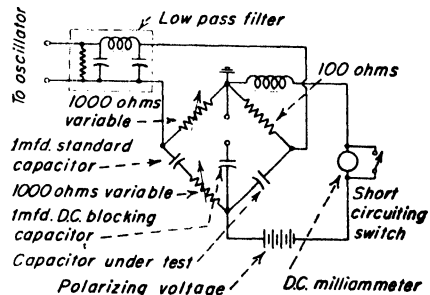



FIG. 15. Capacitance and power-factor measurement.

27. Variable Capacitors. The most common type of variable capacitor consists of a series of parallel metal plates fastened to a shaft capable of rotation so that the moving plates intermesh with a set of fixed plates. Air is the main dielectric in such capacitors, although some solid insulating material is required to ensure that the two sets of plates are correctly located with respect to each other. Many ways of insulating the plates from each other have been devised, using one or more pieces of the insulating material in sheet, rod, or bar form. Phenolic, hard rubber, pyrex, porcelain, fused quartz, and steatite are some of the materials used for such insulators.

The most common use of a variable capacitor is in association with a coil, the combination forming a circuit resonant to a band of radio frequencies depending upon the coil constants and the capacitance range of the capacitor. For a number of applications it is more convenient to have the capacitance change in a different way than proportional to the angle of rotation of the plates. This first resulted in the "decremeter" plate and the straight-line wavelength plate. As the use of frequency rather than wavelength became common, the straight-line frequency plate came into use and later the "mid-line" plate. There are other possibilities such as straight-line percentage wavelength and straight-line percentage frequency, the latter being of advantage in frequency measurements. In any of the above shapes or classifications, the movable plates formerly were so shaped as to give the desired frequency or wavelength curve. This resulted in an ill-shaped plate difficult to balance or to hold to a desired setting. In some cases semicircular rotating plates were used with the fixed plates cut away so as to obtain the desired curve. In any of the special forms of plates the plate shape may vary. The minimum and maximum capacitances of the circuit play a large part in determining the outline of the plate.

Brass or aluminum plates and steel shafts are ordinarily used. If the capacitor is intended for use on high voltages, the spacing between opposite plates must be sufficient to avoid a flashover or arcing between plates. It is customary to round off all sharp edges and corners in such capacitors to avoid flashover.

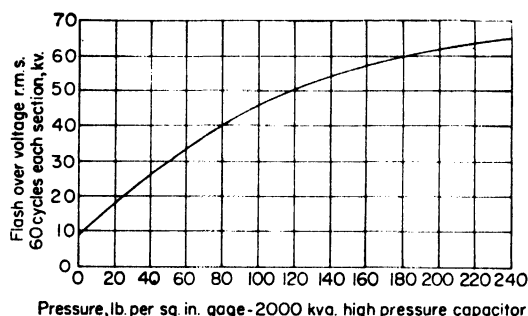



FIG. 16. Flashover voltage at 60 cycles of 2,000-kva capacitor.

Capacitors of the air type are often filled with oil, which increases the voltage that they can stand and increases the capacitance from two to five times depending on the dielectric constant of the oil used.

Compressed-air capacitors were formerly used in some radio transmitting stations. The voltage that such a capacitor will stand is increased without changing the capacitance.

Compressed-gas capacitors, utilizing nitrogen under pressure up to 2,000 psi as the dielectric, are now being extensively used in broadcast transmitters. The advantages of low loss and permanent characteristics have long been recognized, but it is only of recent date that any attempt has been made to offer a wide commercial selection of this type of capacitor.

One manufacturer offers three lines with flashover ratings of 15, 20, and 30 kv rms at 1,000 kc, and capacitance ranges up to 1,000, 1,500, and 2,000 $\mu\mu$f, respectively. These are available in fixed, adjustable, or continuously variable types. Special units have been built with flashovers up to 60 kv rms and capacitances up to 20,000 $\mu\mu$f.

Construction varies somewhat with different manufacturers. One offers a completely nonmagnetic assembly using heat-treated aluminum tank and end closures. As a typical example, a variable capacitor having 30 kv rms flashover rating will have a height of 36 in., an over-all diameter of 12 in., and a weight of 90 lb.

Gases other than nitrogen have been used, some of which show considerable promise in increasing flashover voltage and reducing size and weight. These are available in either fixed or variable capacitance types and in sizes from 100 to 2,000 $\mu\mu$f.

Variable high-voltage vacuum capacitors are among the recent developments. These capacitors are similar to the fixed high-voltage vacuum units but have mechanical arrangements to vary the capacitance a few micromicrofarads in the vicinity of 10 or 20 $\mu\mu$f. A larger unit covers a range from 250 to 350 $\mu\mu$f at 10,000 volts peak. Variable vacuum capacitors in which capacitance varies linearly, with respect to shaft rotation, from 10 to 60 $\mu\mu$f are used as plate tank capacitors in high-frequency applications.

28. Gang Capacitors. The single-dial control radio receiver brought problems to the designer in how to tune two to five circuits accurately using a corresponding number of similar coils and variable capacitors operating on the same shaft. As it is practically impossible conveniently to manufacture two capacitors exactly alike, to say nothing of three or four alike, so that their capacitances shall be exactly the same throughout the complete rotation of the capacitor plates and accurately tune the capacitors with the same number of similar coils which differ slightly in value, it has been customary to balance or equalize these tuned circuits by the addition of small paralleling capacitors called *trimmer* or *padder* capacitors. Such capacitors can be obtained matched to one-half of 1 per cent. It is possible to obtain two to four capacitors called *gang* capacitors for radio receivers arranged with their shafts in line and operated by one dial, matched to one-half of 1 per cent. The individual capacitors may be separated from one another by metal shields if desired.

Earlier forms of trimmer capacitors used a combination of mica and air as the dielectric. Adjustment in capacitance was accomplished by turning a screw which moved one of the plates nearer the other. A recent design incorporates several concentric cylinders supported on a ceramic member and arranged to interweave so as to change the capacitance from 2.5 to 30 $\mu\mu$f. Another trimmer of recent design employs a metalized glass tube with a metal core within, which may be moved with respect to the metalized coating. Several sizes are available with d-c breakdown voltages from 1,500 to 2,000 volts.

A still more recent design employs optically flat base and ceramic rotor. Fired silver electrodes on base or rotor top permit changing capacitance by varying the overlap, but the change per degree of rotation is nearly constant. The temperature coefficient may be controlled within narrow limits.

Fig. 17. Dimensions useful in determining capacitance.

29. Design Equations for Variable Air Capacitors. The capacitance of a capacitor made up of three plates as indicated in Fig. 17 can be obtained by determining the area of the overlapping plates, the distance between the adjacent plates, and substitution of these values in the general equation given in Sec. 12. The area of the shaded portion of Fig. 17 is $\frac{1}{2}\pi(r_1^2 - r_2^2)$. The distance between the plates is $\frac{1}{2}(s - t)$. Substituting these values in the general equation, the capacitance is given by

$$C = \frac{0.0885 \times \frac{1}{2}\pi(r_1^2 - r_2^2) \times (3 - 1)}{\frac{1}{2}(s - t)}$$

The maximum capacitance of a unit with N plates can be obtained by using a similar equation which may be written

$$C = \frac{0.278(r_1^2 - r_2^2)(N - 1)}{(s - t)}$$

In the above equation C is in micromicrofarads and the dimensions r_1, r_2, s, and t in centimeters. These equations neglect the capacitance through the solid insulation which is used in the capacitor and the fringing effect, the correction for which is in

Straight-line wavelength	Straight-line frequency	Straight-line percentage wavelength or frequency
$C_2 = (a_2\theta + b_2)^2$	$C_3 = \dfrac{1}{(a_3\theta + b_3)^2}$	$C_4 = a_4\epsilon^{b_4\theta}$
$A_2 = k\{(a_2\theta + b_2)^2 - \text{resid. cap.}\} + K\theta$	$A_3 = k\left\{\dfrac{1}{a_3\theta + b_3)^2} - \text{resid. cap.}\right\} + K(180 - \theta)$	$A_4 = k\{a_4\epsilon^{b_4\theta} - \text{resid. cap.}\} + K\theta$
$R_2 = \left[114.6\{2ka_2(a_2\theta + b_2) + K\}\right]^{1/2}$	$R_3 = \left[114.6\left\{\dfrac{2ka_3}{(a_3\theta + b_3)^3} + K\right\}\right]^{1/2}$	$R_4 = \left[114.6\{ka_4b_4\epsilon^{b_4\theta} + K\}\right]^{1/2}$
Constants:	Constants:	Constants:
$a_2 = \dfrac{\sqrt{\text{max. cap.}} - \sqrt{\text{resid. cap.}}}{180}$	$a_3 = \dfrac{1}{180}\left\{\dfrac{1}{\sqrt{\text{resid. cap.}}} - b_3\right\}$	$a_4 = \text{resid. cap.}$
$b_2 = \sqrt{\text{resid. cap.}}$	$b_3 = \dfrac{1}{\sqrt{\text{max. cap.}}}$	$b_4 = \dfrac{\text{log. (max. cap)} - \text{log (resid. cap.)}}{78.174}$

Common constants $k = \dfrac{\text{total plate area} - 180\,K}{\text{max. cap.} - \text{resid. cap.}}$

$$K = \frac{r^2}{114.6}$$

Sec. 12. Many capacitors are made to have as small a minimum capacitance as possible, giving a large ratio of maximum to minimum capacitance, but this is of doubtful advantage, as slight changes of capacitance due to warping of plates or wear in bearings will cause a relatively large error at the lower end of the scale but practically no noticeable effect at the maximum capacitance end of the scale.

A semicircular plate capacitor gives a capacitance calibration curve similar to C shown in Fig. 18. With the exception of the portions near the ends of the curve, it is practically a straight line. In practice, the lower 10 and upper 5 or 10 deg of a 180-deg scale are not used, so as to avoid the curvature in the calibration curve in these regions. Zero setting does not give zero capacitance.

A curve for such a capacitor used with a coil is shown by F in Fig. 18. The frequency changes very rapidly on the lower part of the scale. A slight capacitance change would make a large frequency change. Therefore, when using frequency meters having semicircular plate capacitors that constitute the main capacitance of the circuit, the coils should be so designed as to give overlaps without resort to the low-capacitance end of the scale.

Fig. 18. Characteristics of semicircular plate capacitor.

As the wavelength λ of a wavemeter circuit is proportional to $\sqrt{LC}$, if L is assumed to be constant, $\lambda \propto \sqrt{C}$ and $\sqrt{C}$ is proportional to the square root of the setting θ. For a uniform wavelength condenser it is necessary to have C vary as the square of the setting θ, or $C \propto \theta^2$.

θ, deg	Radius, cm		
	R_2	R_3	R_4
0	2.49	8.25	1.93
5	2.56		
10	2.60	6.70	2.02
20	2.76	5.62	2.13
30	2.89	4.80	2.24
40	....	4.17	2.36
60	3.18	3.32	2.64
80	....	2.75	2.98
90	3.56		
100	....	2.37	3.38
120	3.86	2.10	3.85
140	....	1.90	4.40
150	4.12	....	4.71
160	....	1.76	5.04
170	....	....	5.40
180	4.38	1.65	5.80

Again, it may be desirable that the percentage change in capacitance for a given angle of rotation of the plates be the same for all parts of the scale as in the Kolster decremeter.[1] The polar equation for the boundary curve is

$$r = \sqrt{2C_0 a\epsilon^{a\theta} + r_2^2}$$

[1] *Nat. Bur. Standards Circ.*, **74**, p. 117. *Nat. Bur. Standards Sci. Paper* 235.

where C_0 = capacitance when angle $\theta = 0$

a = constant = percentage change of capacitance per scale division

ϵ = 2.71828

r_2 = radius of cutout portion to clear washers separating variable plates

The equations and tables on page 4-20 have been compiled by Griffiths.[1] The three types of plates given are for equivalent capacitors having a capacitance at zero setting *of 36 μμf, with a plate area of 20 sq cm.*

The paper mentioned above gives the data for the radii at different angles for the capacitors mentioned in the table of equations on page 4-21.

30. Effect of Putting Odd-shaped Plate Capacitors in Series or Parallel. If any of the above capacitors are placed in parallel or in series with another capacitor, the straight-line calibration will be altered. If paralleling capacitors are used, the plate shape would require recalculation, after which the plate would become more nearly semicircular. If a capacitor is added in series, the calculation of the plate shape is more difficult. Griffiths[2] gives complete equations for a number of series combinations, the following table applying to the cases indicated where maximum capacitance of variable capacitor = 500 μμf, minimum capacitance of variable capacitor = 36 μμf, series fixed capacitance = 500 μμf, total plate area = 20 sq cm, r = radius of inactive semicircular area of moving plate = 1.2 cm.

θ, deg	Radius, cm			
	R_5	R_6	R_7	R_8
0	2.74	2.16	9.25	1.82
10	2.80	....	6.95	
20	....	2.35	5.57	1.96
30	2.92	....	4.65	
40	....	2.56	....	2.15
50	3.06			
60	....	2.78	3.32	2.38
70	3.22			
90	3.40	....	2.42	2.85
100	....	3.37		
110	3.66			
120	....	....	2.02	3.57
130	3.88			
140	....	4.25		
150	4.18	....	1.78	4.74
160	....	4.85		
170	4.52			
180	4.73	5.66	1.62	7.16

R_5, straight-line capacitance with series fixed capacitance.
R_6, corrected square law of capacitance with series fixed capacitance.
R_7, inverse square law of capacitance with series fixed capacitance.
R_8, exponential law of capacitance with series fixed capacitance.

31. Important Considerations in Design. It is not difficult to find a large number of capacitors on the market that will answer the needs of any application in radio receivers. The manufacture of capacitors for such use has been brought to a high state of development, both electrically and mechanically. The design problems here are simpler in that low power and low voltage are to be handled.

[1] Griffiths, W. H. F., Notes on the Laws of Variable Air Condensers, *Exptl. Wireless and Wireless Engr.*, **3**, 3, January, 1926.

[2] Griffiths, W. H. F., Further Notes on the Laws of Variable Air Condensers, *Exptl. Wireless and Wireless Engr.*, **3**, 743, December, 1926.

When capacitors for radio transmitters are designed, provision must be made for handling high power and high voltage. The use of very high radio frequencies has added to the problem by requiring better insulating materials. Insulators that were satisfactory at low radio frequencies have been found to heat up and be unsuited for frequencies such as 30 to 100 Mc and higher.

The following classification shows how capacitors for transmitting sets could be divided with respect to the voltages to which they are subjected:

Those subjected to steady d-c voltages only.
Those subjected to l-f voltages only.
Those subjected to damped r-f voltages only (obsolete).
Those subjected to steady c-w r-f voltages only.
Those subjected to modulated c-w r-f voltages only.
Those subjected to d-c voltages with superimposed r-f voltage.
Those subjected to l-f voltage and superimposed r-f voltage.

The last four of the above divisions could be further subdivided into those for use on frequencies up to about 3,000 kc, those for use on frequencies from 3,000 to about 25 Mc, and those for use on frequencies of 30 Mc and above. The two latter classes require special construction.

In specifying the rating of capacitors for use in radio transmitters, the following data should be given: capacitance, current, frequency, nature of voltage to be applied. A knowledge of the maximum r-f voltage and maximum current permissible is important. A capacitor should never be operated at more than half the breakdown voltage. In the case of r-f voltages this fraction should be much smaller.

32. Standards of Capacitance. Fixed capacitors using the best grade of mica or fixed air condensers may be used as capacitance standards for audio frequencies. A variable air capacitor is essential as a standard in many radio measurements.

High-grade mica capacitors can be employed as standards after calibration as to capacitance and power factor over the range of frequencies at which they are to be used.

An important requirement of a standard capacitor is that the capacitance remain constant, the prerequisite of which is rigidity of construction, which is more difficult to secure in a variable than in a fixed capacitor. There should be no relative motion possible between the movable plates and the pointer. There should be no stops against which the pointer or movable plates may strike and thus destroy the calibration. The manner of insulating the two sets of plates is of great importance, not only in fulfilling the rigidity requirement, but in minimizing the power loss. An insulating material having a low temperature coefficient of expansion should be used, so that the capacitance will not change perceptibly with temperature. As small an amount of solid insulating material as possible should be employed, keeping it well out of the electric field. This field is quite intense near the high-potential post. All insulation should be avoided in the vicinity of that terminal if power factor is to be kept low.

The capacitor should be provided with a metal shield, which may be grounded during measurements if the capacitance is to remain constant. The leads inside the capacitor should be as short and direct as possible. The resistance of leads, plates, and contacts should be kept to the minimum. Flexible connection to the moving plates should not be used in a capacitor used as a standard.

Although it has been customary and is permissible in some measurements to neglect the small resistance and inductance found in variable air capacitors made for precision laboratory work, yet, as the frequency is increased to 5 Mc and higher, such omissions may result in considerable inaccuracy in the results. These small residual impedances, when taken into account, give an equivalent circuit for the variable air capacitor[1] as shown in Fig. 19, where C is the static capacitance of the capacitor, R the resistance loss in the metal parts, L the inductance of the leads and connections of stacks of plates, and G the conductance or losses in the solid dielectric parts.

[1] Field, R. F., and D. B. Sinclair, A Method for Determining the Residual Inductance and Resistance of a Variable Air Condenser at Radio Frequencies, *Proc. IRE*, **24**, 255, February, 1936.

The resistance R is practically independent of capacitance setting and increases with frequency because of skin effect, being proportional to the square root of the frequency at higher frequencies.

The effect of the inductance L is to cause the apparent capacitance to be greater than the actual capacitance, as may be seen from the following equation:

$$C_a = \frac{C}{1 - \omega^2 LC}$$

where C_a = apparent capacitance, farads
C = actual capacitance, farads
L = inductance of capacitor, henrys
ω = 2π times frequency, cps

The differences between the apparent and the actual capacitances are negligible at medium frequencies but increase rapidly at higher frequencies, where the effect can be reduced by feeding the center of the stack of plates rather than the end. This construction is used in variable air capacitors intended for use as standards or for measurements at the higher radio frequencies.

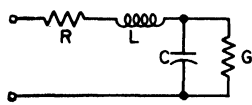



FIG. 19. Equivalent circuit of air capacitor.

The power factor of a standard variable air capacitor at frequencies up to several megacycles is usually much higher at the low portion of the capacitance scale. As the frequency is raised, the power factor increases very rapidly with increase in scale setting.

33. Methods of Measuring Capacitance. There are two general methods of capacitance measurement: (1) absolute measurements in terms of other electrical or physical units, (2) comparison methods, where a capacitor of unknown capacitance is compared with a known calibrated capacitor. The absolute methods are not carried out at radio frequencies.

Among the simpler comparison methods may be mentioned those operating at an audio frequency or low radio frequency where the measuring device may be direct-reading. Such devices would include capacitor analyzers, bridges, and microfarad meters. Some of these devices are intended for quickly indicating capacitance and may not show more than two significant figures. Precision of measurement is sacrificed for speed in use. Such devices are extremely useful in some cases.

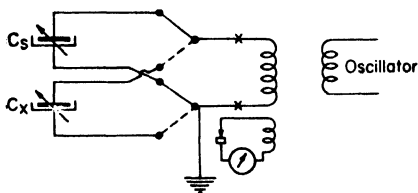



FIG. 20. Circuit for measuring capacitance.

There are a number of arrangements that may be used for capacitor calibrations at radio frequencies. These include the simple radio circuits shown in Figs. 20 and 21, r-f bridges, Q meters, and twin-T impedance-measuring circuits. The precision of measurement in any of the various methods cannot exceed that of the standard capacitor.

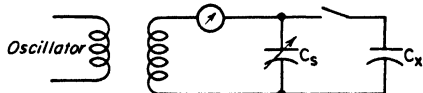



FIG. 21. Simple scheme for measuring capacitance.

A simple tuned circuit consisting of a coil and the capacitor under test is arranged with a double-throw switch so that the standard capacitor may be readily substituted. Resonance may be indicated by a sensitive meter coupled to the main coil by a few turns of wire. A crystal detector and 1-ma d-c meter make a very convenient indicating device. Power is supplied electromagnetically by a small vacuum-tube oscillator. The measurement circuit is shown in Fig. 20. The shielded side of the capacitor should be grounded. It is essential that the leads connecting the switch points to each capacitor be of the same

length in each case; otherwise the circuits will not have the same amount of inductance when one capacitor is substituted for the other. This will result in an error in the calibration. The coupling between the test circuit and the oscillators should be kept quite loose, which will be necessary if a sensitive resonance-indicating instrument is used.

If in the circuit shown in Fig. 20 a fixed inductor is used, the calibration will be made at various frequencies depending upon the capacitance for the different settings. A variable air capacitor of suitable size could be connected across the coil at *XX* and used to keep the resonance frequency the same for any setting of C_x. If such a circuit is carefully set up, no errors will result if the two circuits connected to C_x and C_s are similar. The frequency at which the measurements are made can be measured with a frequency meter. The frequency or frequency range over which a calibration is made should always be stated.

For rougher calibration work, the circuit shown in Fig. 21 may be used where C_s is tuned both with and without C_x in the circuit. It should be noted that the leads and switch connecting C_x to the circuit will introduce errors in the calibration.

The accuracy of this method can be greatly improved by careful arrangement and shortening of the leads connecting C_x to C_s. The switch should be replaced by a short wire, which by the least movement makes or breaks contact with the high-potential terminal of condenser C_x.

A method[1] of precision calibration of variable air capacitors at a single frequency has been described in which the unknown capacitor and the standard capacitor are alternately made a part of the oscillator furnishing the power. The method also offers a very precise means of measuring the change in capacitance with frequency of a mica capacitor.

The manufacturers of equipment for precise measurement of capacitance furnish instructions for use of their instruments. In general, the unknown capacitance is substituted for an equivalent capacitance in the standard capacitor.

The accuracy that can be attained in the measurement of a given capacitor depends upon a number of factors, among which are (1) the type of capacitor, *i.e.*, shielded or unshielded, (2) the length of leads employed in connecting the capacitor, (3) the physical dimensions of the capacitor, which also affect the length of leads, (4) the capacitance of the capacitor. Other factors that may be mentioned are accuracy of standard capacitor, ability of test capacitor to maintain capacitance value, and temperature effects.

A shielded capacitor is the only type that may be expected to maintain its capacitance value in subsequent use. The capacitance of an unshielded unit may readily be measured, but the value so obtained is affected by its position with respect to grounded panels or other objects, by the terminal that is grounded, lead length, etc., so that unshielded capacitors are not satisfactory where precise values of capacitance are required.

The length of leads, the size, and spacing affect the value of capacitance obtained. The leads should be made as short as possible. Corrections can be made for the leads, or their configuration can be given so that they could be duplicated if desired.

A capacitor of large proportions offers some difficulty because of the long connecting leads required for measurement. Above frequencies of 500 kc, lead corrections should be applied, or procedures used in which the leads are the same length for the standard and unknown capacitors.

Very small and very large capacitors are more difficult to measure accurately than those from 10 to 2,000 or 3,000 $\mu\mu$f. The measurement of very small capacitors is discussed in the next section. Very large capacitors are measured by connecting in series with a smaller capacitor, or by paralleling a number of smaller capacitors. In either case there may be sufficient "stray" capacitances to reduce the accuracy of the measurement.

[1] Hall, E. L., and W. D. George, Precision Condenser Calibration at Radio Frequencies, *Electronics*, **7**, 318, 1934.

34. Measurement of Very Small Capacitances. It has always been difficult to obtain agreement between various laboratories in the measurement of capacitances of 5 μμf or smaller. There have been several reasons for this, including differences in the standards of the various laboratories, differences in lead length, size, and spacing, and stray capacitances. Perhaps the greatest difficulty was in the magnitude of the quantity being measured. Consider a capacitance of 1.00 μμf. A 2-in. length of No. 10 wire 0.4 in. above a ground plate would have this capacitance.[1] A measurement to 1 per cent (0.01 μμf) would, therefore, require very exact positioning of leads, or the use of a jig to ensure exact placement of the condenser under test.

Considerable difficulty was experienced by tube manufacturers early in the recent war, in supplying tubes with interelectrode capacitance within the limits prescribed by the Navy. The difficulty was overcome by the adoption of special test equipment and special measurement techniques.

A special bridge[2] has been described for measuring interelectrode capacitances at 465 kc. The conductance and capacitance can be measured in terms of standard capacitors having much larger values than the components being measured. Five decimal divisors are available so that capacitances can be read to 1/100,000 μμf. Bridge balance is sensitive to that amount.

Very small capacitance values or changes in capacitance can be determined by measuring the change in frequency when the unknown capacitance is added to the tuning capacitance of an oscillator. The method described in a previous reference can be adapted by slight rearrangement so as to measure the a-f difference between an auxiliary standard oscillator and the measuring oscillator without the small capacitance and with it in circuit. The small capacitance or change in capacitance is given by the following equation:[3]

$$\Delta C_x = C_0\left[\left(\frac{f_0}{f_1}\right)^2 - 1\right]$$

where ΔC_x = small capacitance, μμf
C_0 = capacitance of oscillator before ΔC_x was added and includes all capacitances that produce the frequency f_0
f_1 = frequency of oscillator with both C_0 and ΔC_x in parallel

The value of C_0 must be measured. This is done in a similar manner by substituting a known capacitor for ΔC_x and measuring the frequency change with an oscillator.

Capacitance differences can be very precisely measured if a standard is available with a sufficiently finely divided scale. Such standards have taken cylindrical form with a concentric rod, movable longitudinally by means of a fine screw and equipped with a scale to indicate the position of the rod. The change in capacitance can be calculated from the following equation:[4]

$$C = \frac{0.613}{\log_{10}(b/a)}$$

where C = capacitance, μμf per in.
b = inside diameter of cylinder, in.
a = diameter of inner rod, in.

35. Methods of Measuring Capacitor Resistance and Power Factor and Dielectric Constant of Insulating Materials at Radio Frequencies. Measurements of capacitor resistance and power factor of insulating materials are made in practically the same manner, as the sample of insulating material is prepared so as to form a capacitor. Methods of measuring capacitor resistance[5] and power factor of insulating materials[6]

[1] Terman, F. E., "Radio Engineers' Handbook," p. 114, McGraw-Hill, 1943.
[2] Young, C. H., Measuring Inter-electrode Capacitances, *Tele. Tech.*, **6**, 68, February, 1947.
[3] Terman, *op. cit.*, p. 924.
[4] *Ibid.*, p. 922.
[5] Radio Instruments and Measurements, *Nat. Bur. Standards Circ.*, **74**, p. 190.
[6] Methods of Measurement of Properties of Electrical Insulating Materials, *Nat. Bur. Standards Sci. Paper* 471.

have been given in publications of the National Bureau of Standards. The ASTM has several standard methods of testing electrical insulating materials for power factor and dielectric constant.[1]

The circuit shown in Fig. 22 may be used for measurements of resistance, power factor, and dielectric constant. Assuming that the power factor of a sample of insulating material is to be measured, the sample in sheet form is made into a capacitor of capacitance between 100 and 1,000 $\mu\mu$f, as represented by C_x (Fig. 22). The remainder of the circuit consists of the coil L, thermoelement T, and double-pole, double-throw switch S, in which resistors R may be inserted. The galvanometer G gives deflections

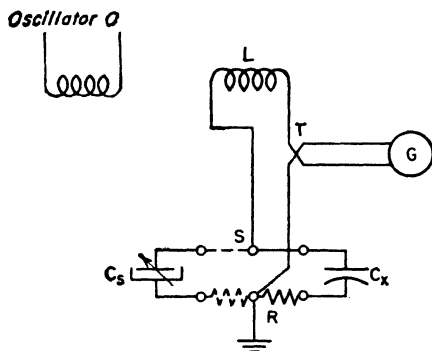



FIG. 22. Circuit for measuring properties of insulators.

that are proportional to the square of the current flowing in the circuit LTC_xR, as electromagnetically induced from the r-f oscillator O.

The deflections of galvanometer G are noted for several values of inserted resistance R and for the case when R is a link of practically zero resistance. Using the "zero resistance" deflection and the deflection for a known value r of resistance inserted in switch S, the resistance R_T of the total circuit LTC_xR is given by

$$R_T = \frac{r}{\sqrt{\frac{d_0}{d_1}} - 1}$$

The average of the values of R_T calculated for various values of r should be taken as the resistance of the complete circuit. The resistance R_S of the circuit when C_S is substituted for C_X should be obtained in the same manner. The resistance R_X of the capacitor C_x is then given by $R_x = R_T - R_S$. It is essential for this measurement that the two parts of the circuit which are interchanged should be as nearly identical as possible.

After the resistance R_x of the insulating material capacitor is obtained, the power factor or phase difference can be calculated from the equations given above. The dielectric constant K can be calculated from the equation $K = Ct/0.0885S$, where C = capacitance of sample in micromicrofarads, t = thickness of sample in centimeters, and S = area of smaller plate in square centimeters. The capacitance is known, as given by C_S, and the area of one plate and the thickness of the sample can easily be measured.

The method described above operates satisfactorily at frequencies from 100 to 1,500 kc.

ASTM method D150-47T gives details for the procedure for eight methods of

[1] Tentative Methods of Test for Power Factor and Dielectric Constant of Electrical Insulating Materials, designation D150-47T; Standard Method of Test for Power Factor and Dielectric Constant of Natural Mica, designation D1082-51T.

measuring dielectric constant and power factor of solid and liquid electrical insulating materials. These include five bridge techniques and three resonance techniques which are applicable in the range of about 25 cps to a few hundred megacycles.

Transmission-line and concentric-line methods of impedance measurement can be adapted to power factor and dielectric constant measurements of dielectrics, but there are practical difficulties in the preparation of samples which limit the usefulness of these methods. A resonant-cavity method[1] has been described operating in the range from 100 to 1,000 Mc which is claimed to have many advantages in use, such as simplicity and rapidity. Several cavities are required, each covering a frequency ratio of 1 to 1.5. Dielectric samples in disk form up to 1½ in. in diameter and from 0.05 to ¼ in. thick are used. From measurements made with the sample in and out of the cavity, the dissipation factor and the dielectric constant are obtained by simple calculation.

[1] Works, C. N., T. W. Dakin, and F. W. Boggs, A Resonant-cavity Method for Measuring Dielectric Properties of Ultrahigh Frequencies, *Trans. AIEE*, **63**, 1092, 1944; *Proc. IRE*, **33**, 245, April, 1945.

CHAPTER 5

COMBINED CIRCUITS OF L, C, AND R

By W. F. Lanterman[1]

1. Transient and Steady-state Currents. When a voltage is suddenly applied to a circuit, the current assumes a transient state for a brief interval, then gradually settles down to a steady-state condition which it maintains until the voltage is interrupted or changed. Relations for computing transient and steady-state currents and voltages in LCR circuits are given in the following paragraphs. The curves shown for current and voltage are typical for each case but are drawn for particular values of the arbitrary constants.

TRANSIENT CURRENTS IN LCR CIRCUITS

2. Symbols Used in Transient Expressions. In the transient exrpessions given in Secs. 3 to 5, the following symbols will be used:

L = inductance, henrys
C = capacitance, farads
R = resistance, ohms
T = time constant, sec; time, sec, for current or voltage to reach $1/\epsilon$ or approximately 36.8 per cent of its initial value if decreasing; or $1 - (1/\epsilon)$, or approximately 63.2 per cent of its final value if increasing
i = instantaneous current, amp, at time t
e = instantaneous voltage, volts, at time t
e_c = instantaneous capacitor voltage
e_L = instantaneous inductor voltage
t = time, sec, after starting
I = steady-state d.c., amp
E = maximum value of a-c voltage, volts
V = steady-state d-c voltage, volts
Q = capacitor charge, coulombs
Z = a-c impedance, ohms
$= \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$ for LCR circuit
$= \sqrt{R^2 + (\omega L)^2}$ for LR circuit
$= \sqrt{R^2 + (1/\omega C)^2}$ for RC circuit
f = frequency of applied a.c., cps
ω = angular velocity of applied a.c., radians per sec $= 2\pi f$
f_1 = natural frequency of oscillatory circuit LCR, cps
ω_1 = natural angular velocity of oscillatory circuit LCR, radians per sec $= 2\pi f_1$
W_R = energy, joules, dissipated in R during transient state
W_L = energy, joules, stored by or lost by L during transient state
W_C = energy, joules, stored by or lost by C during transient state
θ = phase angle of a-c voltage at $t = 0$, *i.e.*, when switch is closed
ϕ = phase angle of impedance as defined for each case
ϵ = 2.718 (base of natural logarithms)
α and β = (defined in Sec. 5)

[1] National Broadcasting Co., Inc., Chicago, Ill.

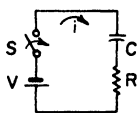
3. *RC* Circuit Transients.

NOTE: The following formulas for i in RC circuits are not true for extremely small values of t. For very small t, the L of the circuit, no matter how small, limits i, and the relations of Sec. 5 for LCR circuits must be applied. This is especially important for short pulses or high frequencies, where small values of t are involved. The equations for RC circuits neglect the effects of L for small values of t, but the figures show true current forms.

1. *D-c Voltage V Suddenly Applied to Deenergized RC.*

FIG. 1.

$$i = \frac{V}{R}\epsilon^{-(t/RC)}$$

$$e_c{}^* = V[\underline{1} - \epsilon^{-(t/RC)}]$$

$$e_R = V\epsilon^{-(t/RC)}$$

$$T = RC$$

$$W_R = W_C = \tfrac{1}{2}CV^2 = \tfrac{1}{2}\frac{Q^2}{C}$$

FIG. 2.

For fully charged C ($t = \infty$),

$$W_C = W_R = \tfrac{1}{2}QV$$

2. *C Charged to Voltage V and Suddenly Discharged through R.*

FIG. 3.

$$i = -\frac{V}{R}\epsilon^{-(t/RC)}$$

$$e_c = V\epsilon^{-(t/RC)}$$

$$e_R = -V\epsilon^{-(t/RC)}$$

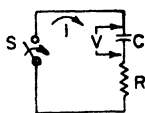
$$T = RC$$

$$W_C = W_R = \tfrac{1}{2}CV^2 = \tfrac{1}{2}\frac{Q^2}{C}$$

Time

Voltage or current

i, eR & -ec

FIG. 4.

For complete discharge of C ($t = \infty$),

$$W_R = \tfrac{1}{2}QV$$

3. *A-c Voltage e Suddenly Applied to Deenergized RC.*

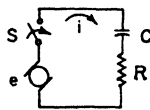



FIG. 5.

Applied voltage $e = E \sin(\omega t + \theta)$

FIG. 6.

$$i^* = \underline{\frac{E}{Z}\sin(\omega t + \theta + \phi)} - \frac{E}{Z\omega RC}\cos(\theta + \phi)\,\epsilon^{-(t/RC)}$$

$$e_C{}^* = \underline{-\frac{E}{Z\omega C}\cos(\omega t + \theta + \phi)} + \frac{E}{Z\omega C}\cos(\theta + \phi)\,\epsilon^{-(t/RC)}$$

$$e_R{}^* = \underline{\frac{ER}{Z}\sin(\omega t + \theta + \phi)} - \frac{E}{Z\omega C}\cos(\theta + \phi)\,\epsilon^{-(t/RC)}$$

Magnitude of $\phi = \cot^{-1}\omega RC$

NOTE: Negative sense of ϕ has been taken into account in above equations. In substituting, use + value of magnitude for ϕ.

There is no transient if $(\theta + \phi) = 90$ or $270°$.

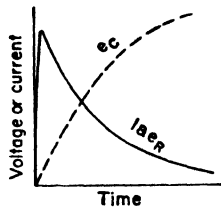
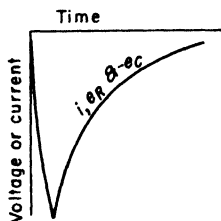
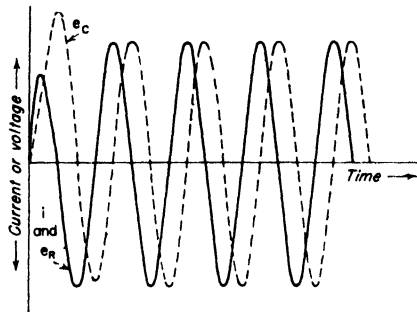
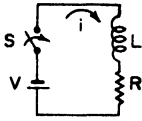
* Underscored terms represent steady-state values; remaining term or terms are the transients.

4. *LR* Circuit Transients. 1. *D-c Voltage V Suddenly Applied to Deenergized LR.*

FIG. 7.

$$i^* = \frac{V}{R}[\underline{1} - \epsilon^{-(Rt/L)}]$$

$$e_L = V\epsilon^{-(Rt/L)}$$

$$e_R^* = V[\underline{1} - \epsilon^{-(Rt/L)}]$$

$$T = \frac{L}{R}$$

$$W_L = \tfrac{1}{2}LI^2$$

For $t = \infty$,

$$W_L = \tfrac{1}{2}\frac{LV^2}{R^2}$$

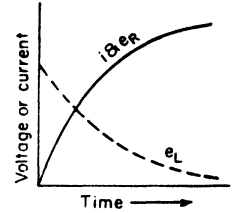



FIG. 8.

2. *LR Carrying Steady D-c Suddenly Short-circuited.*

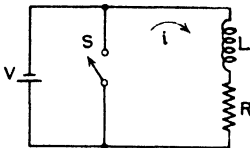



FIG. 9.

$$\text{Steady current } I = \frac{V}{R}$$

$$i = \frac{V}{R}\epsilon^{-(Rt/L)} = I\epsilon^{-(Rt/L)}$$

$$e_L = -IR\epsilon^{-(Rt/L)}$$

$$e_R = IR\epsilon^{-(Rt/L)}$$

$$T = \frac{L}{R}$$

$$W_L = \tfrac{1}{2}LI^2$$

For $t = \infty$,

$$W_L = \tfrac{1}{2}\frac{LV_R^2}{R^2}$$

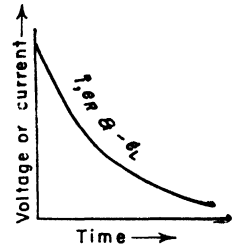



FIG. 10.

3. *A-c Voltage e Suddenly Applied to Deenergized LR.*

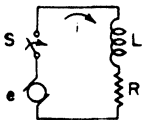



FIG. 11.

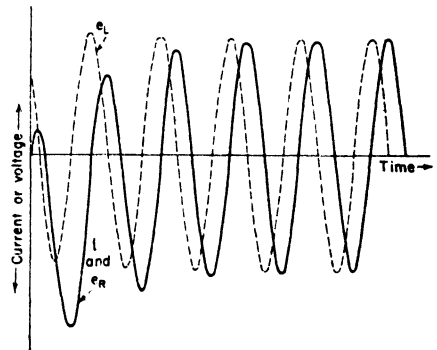
Applied voltage $e = E \sin(\omega t + \theta)$

FIG. 12.

$$i^* = \underline{\frac{E}{Z}\sin(\omega t + \theta - \phi)} - \frac{E}{Z}\sin(\theta - \phi)\epsilon^{-(Rt/L)}$$

$$e_L^* = \underline{\frac{E}{Z}\omega L\cos(\omega t + \theta - \phi)} + \frac{E}{Z}R\sin(\theta - \phi)\epsilon^{-(Rt/L)}$$

$$e_R^* = \underline{\frac{ER}{Z}\sin(\omega t + \theta - \phi)} - \frac{ER}{Z}\sin(\theta - \phi)\epsilon^{-(Rt/L)}$$

$$\text{Magnitude of } \phi = \tan^{-1}\left(\frac{\omega L}{R}\right)$$

There is no transient if $\theta = \phi$ or if $(\theta - \phi) = 180°$.

* Underscored terms represent steady-state values; remaining term or terms are the transients.

5. *LCR* Circuit Transients. 1. *D-c Voltage V Suddenly Applied to Deenergized LCR. General Solutions:*

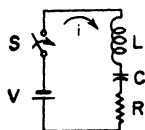



Fig. 13.

$$i = \frac{V}{2\beta L}\epsilon^{-\alpha t}(\epsilon^{\beta t} - \epsilon^{-\beta t}) = \frac{V}{\beta L}\epsilon^{-\alpha t}\sinh \beta t$$

$$e_c{}^* = V\left[\underline{1} - \epsilon^{-\alpha t}\left(\cosh \beta t + \frac{\alpha}{\beta}\sinh \beta t\right)\right]$$

$$e_L = V\epsilon^{-\alpha t}\left(\cosh \beta t - \frac{\alpha}{\beta}\sinh \beta t\right)$$

$$e_R = 2V\frac{\alpha}{\beta}\epsilon^{-\alpha t}\sinh \beta t$$

where

$$\alpha = \frac{R}{2L}$$

$$\beta = \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} = \frac{1}{2L}\sqrt{R^2 - \frac{4L}{C}}$$

There are three special cases, depending upon whether R^2 is greater than, less than, or equal to $4L/C$:

Case I. Aperiodic current, when $R^2 > 4L/C$ (β is real). i, e_c, e_L, and e_R are determined directly from the above general solutions by numerical substitution for α and β.

At $t_1 = \left(\frac{1}{2\beta}\log_\epsilon \frac{\alpha + \beta}{\alpha - \beta}\right)$ sec $= \left(\frac{1}{\beta}\tanh^{-1}\frac{\beta}{\alpha}\right)$ sec, i and e_R reach maxima, then slowly decay to zero. At the same time, t_1, e_L becomes zero and thereafter becomes negative, reaching a minimum at $t_2 = \left(\frac{1}{\beta}\log_\epsilon \frac{\alpha + \beta}{\alpha - \beta}\right)$ sec $= \left(\frac{2}{\beta}\tan^{-1}\frac{\beta}{\alpha}\right)$ sec, which is twice the time required for i to reach maximum.

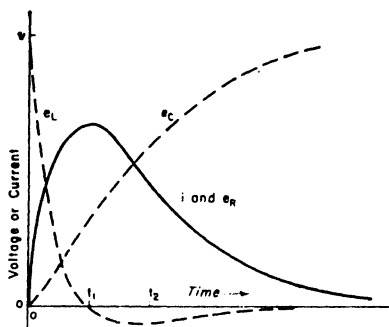



Fig. 14. Case I. Aperiodic.

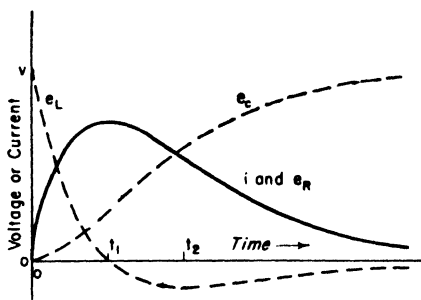



Fig. 15. Case II. Critical damping

Case II. Critical damping, when $R^2 = 4L/C$ ($\beta = 0$).

$$i = \frac{V}{L}t\epsilon^{-\alpha t}$$

$$e_c{}^* = V[\underline{1} - \epsilon^{-\alpha t}(1 + \alpha t)]$$

$$e_L = V\epsilon^{-\alpha t}(1 - \alpha t)$$

$$e_R = 2\,V\alpha t\epsilon^{-\alpha t}$$

At $t_1 = 1/\alpha = 2L/R$ sec, i reaches a maximum of $0.736V/R$ and e_R reaches a maximum of $0.736V$. At the same time, t_1, e_L becomes zero and thereafter becomes negative, reaching a minimum of $-0.1353V$ at $t_2 = 2/\alpha$ sec, which is twice t_1.

* Underscored terms represent steady-state values; remaining term or terms are the transients.

Case III. Oscillatory current, when $R^2 < 4L/C$ (β is imaginary).

$$i = \frac{V}{\omega_1 L}\epsilon^{-\alpha t}\sin\omega_1 t$$

where $\omega_1 = \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}} = 2\pi f_1$

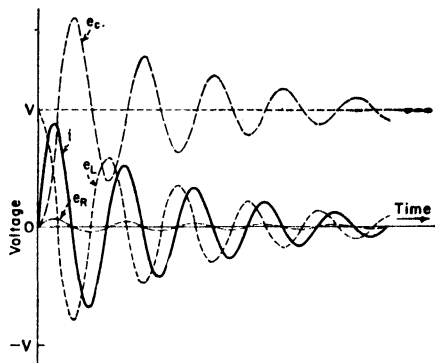
$$e_c{}^* = V\left[\underline{1} - \epsilon^{-\alpha t}\left(\cos\omega_1 t + \frac{\alpha}{\omega_1}\sin\omega_1 t\right)\right]$$

$$e_L = V\epsilon^{-\alpha t}\left(\cos\omega_1 t - \frac{\alpha}{\omega_1}\sin\omega_1 t\right)$$

$$e_R = 2V\epsilon^{-\alpha t}\frac{\alpha}{\omega_1}\sin\omega_1 t$$

NOTE: $\frac{\alpha}{\omega_1} = \sqrt{\frac{R^2C}{4L - R^2C}}$.

FIG. 16. Case III. Oscillatory.

At $t_1 = 1/4f_1$ sec, approximately, i reaches a maximum and at $t_2 = 1/2f_1$ sec, e_c reaches a maximum. e_L is maximum at $t = 0$. Beyond their respective maxima all four quantities oscillate at frequency f_1, with amplitudes slowly decreasing to zero, except e_c, which approaches V.

Superimposed Curves. Curves for i in the three special cases are shown superimposed for direct comparison in Fig. 17. The conditions for all three cases are identical except that R is varied to alter the value of α.

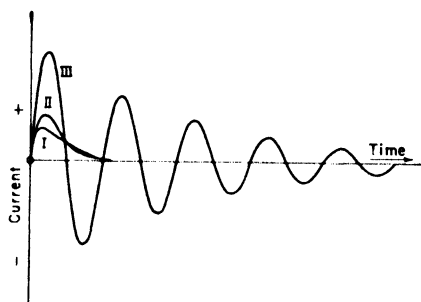



FIG. 17. Superimposed current curves for d.c. in LCR circuit.

2. *A-c Voltage e Suddenly Applied to Deenergized LCR.*

Applied voltage $e = E\sin(\omega t + \theta)$

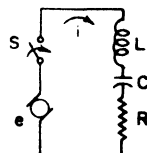



FIG. 18.

There are three special cases, depending upon the ratio

$$\alpha = R/2L$$

Case I. Aperiodic current, when $\alpha^2 = R^2/4L^2 > 1/LC$ (β is real).

$$i^* = \underline{\frac{E}{Z}\sin(\omega t + \theta - \phi)} + \frac{E}{Z}\frac{1}{2\beta L}\left[L(\alpha - \beta)\sin(\theta - \phi) - \frac{1}{\omega C}\cos(\theta - \phi)\right]\epsilon^{-(\alpha-\beta)t} - \frac{E}{Z}\frac{1}{2\beta L}\left[L(\alpha + \beta)\sin(\theta - \phi) - \frac{1}{\omega C}\cos(\theta - \phi)\right]\epsilon^{-(\alpha+\beta)t}$$

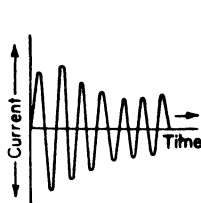



FIG. 19. Case I. Aperiodic.

* Underscored terms represent steady-state values; remaining term or terms are transients.

Case II. Critical damping, when $\alpha^2 \equiv R^2/4L^2 = 1/LC$ ($\beta = 0$).

$$i^* = \underline{\frac{E}{Z}\sin(\omega t + \theta - \phi)} + \frac{E}{Z}\left[(\alpha t - 1)\sin(\theta - \phi) - \frac{1}{L\omega C}\cos(\theta - \phi)\right]\epsilon^{-\alpha t}$$

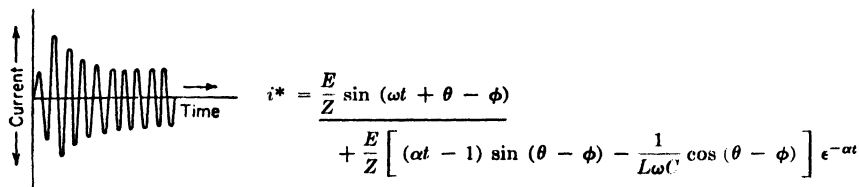



FIG. 20. Case II. Critical damping.

Case III. Oscillatory current, when $\alpha^2 = R^2/4L^2 < 1/LC$ (β is imaginary).

$$i^* = \underline{\frac{E}{Z}\sin(\omega t + \theta - \phi)} - \frac{E}{Z}\epsilon^{-\alpha t}\left[\sin(\theta - \phi)\cos\omega_1 t + \frac{1}{\omega_1 L\omega C}\cos(\theta - \phi)\sin\omega_1 t - \frac{R}{2\omega_1 L}\sin(\theta - \phi)\sin\omega_1 t\right]$$

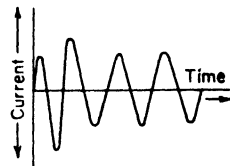



FIG. 21. Case III. Oscillatory.

6. Time Constants. Time constant is an arbitrary factor used to gage the exponential rate of transient current or voltage increase in a circuit containing reactance and resistance in series. It denotes only the relative duration of the transient, and no special physical phenomenon occurs when the transient has existed for a length of time equal to the time constant.

For exponentially increasing quantities the time constant is the time in seconds required for the quantity to attain a value equal to $1 - (1/\epsilon) = 0.63212$ (or 63.2 per cent, approximately) of its final value. For exponentially decreasing quantities, it is the time in seconds required for the quantity to attain $1/\epsilon = 0.36788$ (or 36.8 per cent, approximately) of its initial value.

In series LR circuits, the time constant is $T = L/R$ sec; in series RC circuits, $T = RC$ sec, where R is in ohms, L in henrys, and C in farads (see Secs. 3 and 4).

7. Logarithmic Decrement. The logarithmic decrement is the natural logarithm of the ratio of amplitudes of two successive oscillations (of the same polarity) of a damped wave train. In terms of L, R, C, and Q, the decrement is

$$\delta = \frac{2\pi R}{\sqrt{4(L/C) - R^2}} = \frac{R}{2Lf_r} = \frac{\pi}{Q} \quad (1)$$

where f_r is the resonant frequency of the circuit, and $Q = \omega_r L/R$.

FIG. 22. Typical networks used with pulses.

8. Pulsed Circuits. In television synchronizing systems and other timing circuits the input voltage applied to networks of the types shown in Fig. 22 consists of a series of one or more d-c pulses of the forms shown in Fig. 23. Idealized, the pulse voltage is zero up to a given instant and, at each instant thereafter up to time $= P$ sec, it has a succession of finite values determined by some given pulse form. At P sec, the voltage drops instantly to zero and remains equal to zero until the start of the next pulse. The networks and pulses in Figs. 22 and 23 are typical only; any

* Underscored terms represent steady-state values; remaining term or terms are the transients.

one of an infinite number of pulse shapes may be applied to any one of an infinite number of networks.

Solutions for the output voltage or current are obtained by considering the pulse as creating a transient condition starting at $t = 0$, followed by a new transient condition (to restore the input voltage to zero) starting at $t = P$. If the pulse form can be expressed as a function of t, $e = f(t)$, the output voltage or current is obtained by setting up and solving the differential equations in which the sum of the instantaneous voltage drops in the circuit is equated to $f(t)$. At $t = P$ there is a discontinuity where e suddenly decreases to zero, if the pulse shape is not such that e is already zero when $t = P$. This requires solution of a new differential equation in which the voltage drops existing in the circuit at $t = P$ are equated to zero. This equation must take into account the energy stored in the inductances and capacitances at the instant when $t = P$. Unfortunately, each pulse circuit and each pulse shape is a special case, and results cannot be generally formulated except for typical examples.

FIG. 23. Typical pulse forms.

Example: For a saw-tooth pulse of the form of Fig. 23*b* applied to the *LR* network of Fig. 22*d* the differential equation applying between $t = 0$ and $t = P$ is

$$L\frac{di}{dt} + iR = Vat \tag{2}$$

in which $i = 0$ when $t = 0$, and V and a are defined in Fig. 25. The solution is

$$i = \frac{Vat}{R} - \frac{VaL}{R^2}[1 - \epsilon^{-(Rt/L)}] \tag{3}$$

and the output voltage is

$$e_L = L\frac{di}{dt} = \frac{VaL}{R}[1 - \epsilon^{-(Rt/L)}] = V_n[1 - \epsilon^{-(t/nP)}] \tag{4}$$

where symbols are as defined in Fig. 25.

At $t = P$ the circuit is short-circuited and corresponds to an *LR* circuit carrying d-c I and suddenly short-circuited as in Sec. 4. The voltage equation for $t > P$ is

$$L\frac{di}{d(t-P)} + iR = 0 \tag{5}$$

in which $i = I$ when $t = P$. The solution is

$$i = I\epsilon^{-[R(t-P)/L]} \tag{6}$$

and the output voltage is

$$e_L = L\frac{di}{d(t-P)} = -IR\epsilon^{-[R(t-P)/L]} = -V[1 - n(1 - \epsilon^{-1/n})]\epsilon^{-(t-P)nP} \tag{7}$$

In any practical case, the voltage drop in the source and the effect of the load upon the pulse output cannot be ignored. That is, the L's, C's, and R's in the source and load must be considered as parts of the network. To preserve the pulse shape through a network, it is customary to employ a low-impedance source and a high-impedance load.

Figures 24 to 27 show output voltages for the four pulse forms of Fig. 23 applied to

the networks of Fig. 22. In this case, it is assumed that the input is short-circuited at $t = P$ and thereafter, also that the load impedance is very large compared to that of the network element from which the output voltage is taken. It is also assumed that the Q's of L and C are sufficiently large so that their resistances are negligible as compared to the R's of the network.

Symbols and relations used:
t = time, sec, from start of pulse
T = time constant, sec
P = pulse duration, sec
V = maximum pulse voltage (input)
n = T/P (or $T = nP$)

V e = V (t<P)
O time → P
Applied voltage pulse

Fig. 24. Equations and pulse forms for output voltage of pulsed networks with rectangular pulses of duration P sec applied.

The exponential pulse forms are chiefly of interest when the origin of the input pulse to a particular network is a preceding network whose output pulse has the exponential form. For example, the output pulse is exponential when a rectangular pulse is applied to any of the four networks of Fig. 22.

9. Differentiating and Integrating Networks. Circuits of the types of Fig. 22*c* and *d*, in which the time constant is made very small as compared to the pulse length,

produce output voltage pulses that are approximately the derivatives of the input voltage pulses, as shown in Fig. 28. From this fact, they derive the name *differentiating networks.*

Similarly, circuits of the types of Fig. 22*a* and *b*, in which the time constant is made

Circuit		e = (series C, shunt R)	e = (series R, shunt L)	e = (series R, shunt C)	e = (series L, shunt R)
Time constant		RC	L/R	RC	L/R
e_{out}	$t<P$	$Vn(1-\epsilon^{-\frac{t}{nP}})$		$V\left[\frac{t}{P}-n(1-\epsilon^{-\frac{t}{nP}})\right]$	
	$t=P$	$Vn(1-\epsilon^{-\frac{1}{n}})$		$V\left[1-n(1-\epsilon^{-\frac{1}{n}})\right]$	
	$t>P$	$-V\left[1-n(1-\epsilon^{-\frac{1}{n}})\right]\epsilon^{-\frac{t-P}{nP}}$		$V\left[1-n(1-\epsilon^{-\frac{1}{n}})\right]\epsilon^{-\frac{t-P}{nP}}$	
e_{out}	$T \ll P$, $n \to 0$	V- ; O t → P (Differentiating)		V- ; O t → P	
	$T = P$, $n = 1$	V- ; O t → P		V- ; O t → P	
	$T \gg P$, $n \to \infty$	V- ; O t → P		V- ; O t → P (Integrating)	

Symbols and relations used:
- t = time, sec, from start of pulse
- T = time constant, sec
- P = pulse duration, sec
- V = maximum pulse voltage (input)
- n = T/P (or $T = nP$)
- a = arbitrary constant

V- ; O ; P ; time →

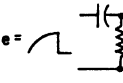
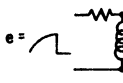
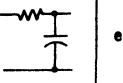
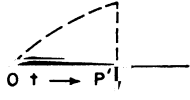
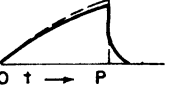
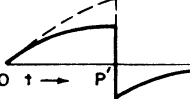
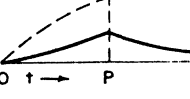
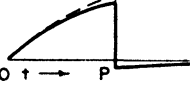
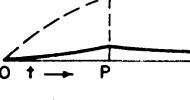
$e = Vat_{(t<P)}$ $V = aP$

Applied voltage pulse

FIG. 25. Equations and pulse forms for output voltage of pulsed networks with saw-tooth pulses of duration P sec applied.

large compared to the pulse duration, produce output voltage pulses that are approximately the integrals of the input voltage pulses, as shown in Fig. 29. This type of network is called an *integrating network.*

10. "Constant-current" RC Circuit. In a series RC circuit as shown in Fig. 30*a*, when C is discharged through R, the current varies logarithmically with time (see

Fig. 4, Sec. 3). For pulses of short duration relative to the time constant R_0C an approximation to constant current is obtained by inserting the network L-R as in Fig. 30*b*, where the time constant L/R is made much larger than the pulse duration time. L and R are determined by the relation $L = R^2C$, and the resulting current is

FIG. 26. Equations and pulse forms for output voltage of pulsed networks with exponentially increasing pulses of duration P sec applied.

essentially independent of the load resistance R_0 except in magnitude. The actual variation of the value of i during such a pulse is reduced to about 10 per cent of the variation obtained with the simple R_0C circuit.[1]

[1] Fundingsland, O. T., and G. J. Wheeler, *Electronics*, November, 1946, p. 130.

Symbols and relations used:

t = time, sec, from start of pulse
T = time constant, sec
P = pulse duration, sec
V = maximum pulse voltage (input)
n = T/P (or $T = nP$)
a = arbitrary constant
α = aT

V e = $V\epsilon^{-at}$ (t<P)
0 time→ P
Applied voltage pulse

FIG. 27. Equations and pulse forms for output voltage of pulsed networks with exponentially decreasing pulses of duration P sec applied.

STEADY-STATE CURRENTS IN *LCR* CIRCUITS

11. *Q* in *LCR* Circuits. By definition,

$$Q = \frac{\text{volt-amp stored}}{\text{watts dissipated}} = \frac{X}{R} \tag{8}$$

Q **of a coil:**

$$Q = \frac{\omega L}{R} = \frac{2\pi f L}{R} \tag{9}$$

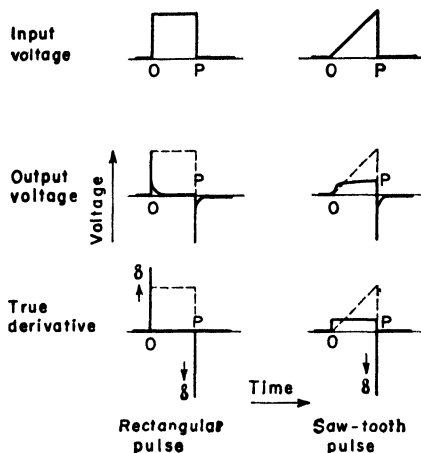



FIG. 28. Pulse forms produced by differentiating networks.

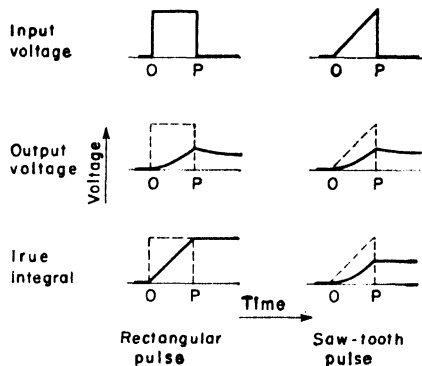



FIG. 29. Pulse forms produced by integrating networks.

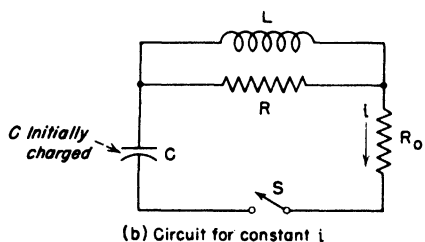
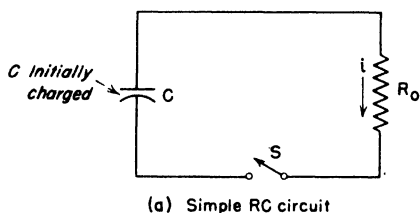



FIG. 30. Modified RC circuit for short pulses of constant current from discharge of C.

Q of a capacitor:

$$Q = \frac{1}{\omega CR} = \frac{1}{2\pi fCR} \tag{10}$$

Q of a circuit:

$$Q = \frac{\omega L}{R_t} \quad \text{or} \quad \frac{1}{\omega CR_t} \tag{11}$$

where R_t is the total resistance and L or C is the total inductance or capacitance in the circuit. If both L and C are present, $\omega = 2\pi \times$ the resonance frequency.

As a rule, the Q of a coil or capacitor varies less with frequency than does the effective resistance. Thus the assumption that Q is constant is usually more nearly accurate than the assumption that R is constant in expressions for tuned circuits. Expressions containing Q are still approximations, however, and are often loosely used without regard for the fact that Q actually does vary with frequency. Such expressions are most nearly correct for large values of Q (of the order of 100) and when the frequency range is small.

In general, the "sharpness" of resonance of a circuit increases with increased values of Q.

12. A-c Impedance Relations. For applied voltage, $e = E \sin \omega t$, $\omega = 2\pi f$, f = frequency in cycles.

$$Z = \frac{E}{I} = R + jX = |Z| \,\underline{/\phi} \quad \text{ohms} \tag{12}$$

$$Z = \sqrt{R^2 + X^2} \quad \text{ohms} \tag{13}$$

$$\phi = \tan^{-1}\frac{X}{R} = \cos^{-1}\frac{R}{Z} = \sin^{-1}\frac{X}{Z} \quad \text{radians or deg} \tag{14}$$

FIG. 31. Equivalent impedances of series combinations of *L*, *C*, and *R*.

Reactance of C:

$$X_C = -\frac{1}{\omega C} = -\frac{1}{2\pi f C} \quad \text{ohms} \tag{15}$$

Reactance of L:

$$X_L = \omega L = 2\pi f L \quad \text{ohms} \tag{16}$$

Impedance of LCR in series:

$$Z = R + j\left(\omega L - \frac{1}{\omega C}\right) \quad \text{ohms} \tag{17}$$

Impedance in series:

$$Z_0 = Z_1 + Z_2 + Z_3 + \cdots \qquad \text{(see Fig. 31)} \tag{18}$$

Two impedances in parallel:

$$Z_0 = \frac{Z_1 Z_2}{Z_1 + Z_2} \tag{19}$$

Three impedances in parallel:

$$Z_0 = \frac{Z_1 Z_2 Z_3}{Z_1 Z_2 + Z_1 Z_3 + Z_2 Z_3} \tag{19a}$$

Any number n of impedances in parallel:

$$Z_0 = \frac{1}{\dfrac{1}{Z_1} + \dfrac{1}{Z_2} + \dfrac{1}{Z_3} + \cdots + \dfrac{1}{Z_n}} \tag{19b}$$

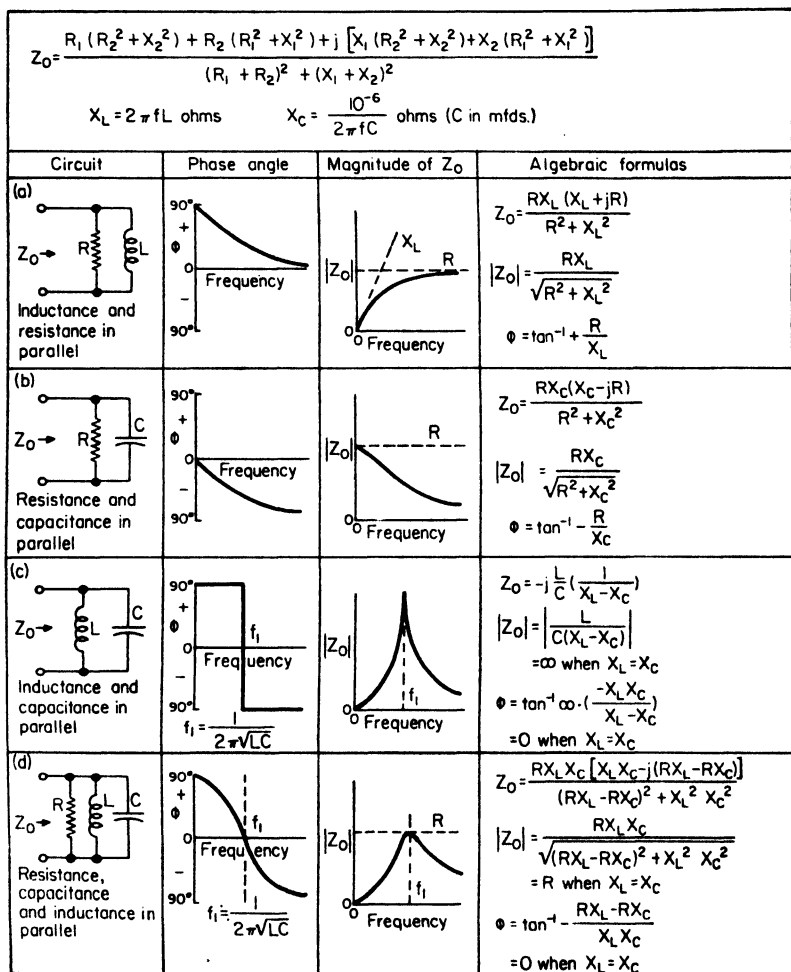
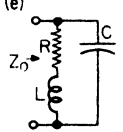
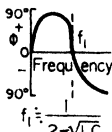
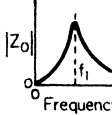
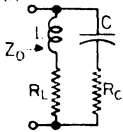
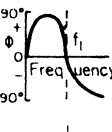
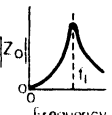



FIG. 32a. Equivalent impedances of parallel combinations of L, C, and R.

$$Z_0 = \frac{R_1(R_2^2+X_2^2)+R_2(R_1^2+X_1^2)+j\left[X_1(R_2^2+X_2^2)+X_2(R_1^2+X_1^2)\right]}{(R_1+R_2)^2+(X_1+X_2)^2}$$

$X_L = 2\pi fL$ ohms $\qquad X_C = \frac{10^6}{2\pi fC}$ ohms (C in mfds.)

Circuit	Phase angle	Magnitude of Z_0	Algebraic formulas
(e) Resistance and inductance in series and in parallel with capacitance	$f_1 \doteq \frac{1}{2\pi\sqrt{LC}}$ Max. + Φ angle occurs when $f = 0.577f_1, Q > 1$ Not valid for $Q \leqq 1$		$Z_0 = \frac{RX_C^2 - j\left[\frac{L}{C}(X_L - X_C) + R^2X_C\right]}{R^2+(X_L-X_C)^2}$ $\lvert Z_0\rvert = X_C\sqrt{\frac{R^2+X_L^2}{R^2+(X_L-X_C)^2}}$ $\doteq \frac{L}{RC}$ when $X_L = X_C$ and R is small $\Phi = \tan^{-1}\left(-\frac{X_L(X_L-X_C)+R^2}{RX_C}\right)$ $\doteq 0$ when $X_L = X_C$ and R is small
(f) Resistance and inductance in series and in parallel with capacitance and resistance in series	$f_1 \doteq \frac{1}{2\pi\sqrt{LC}}$ $Q > 1$ Not valid for $Q \leqq 1$		$Z_0 = \frac{(R_LA+R_CB)+j(X_LA-X_CB)}{(R_L+R_C)^2+(X_L-X_C)^2}$ $A = (R_C^2+X_C^2)$, $B = (R_L^2+X_L^2)$ $\lvert Z_0\rvert = \frac{\sqrt{(R_LA+R_CB)^2+(X_LA-X_CB)^2}}{(R_L+R_C)^2+(X_L-X_C)^2}$ $\doteq \frac{L}{(R_L+R_C)C}$ when $X_L = X_C$ and R_L and R_C are small $\Phi = \tan^{-1}\frac{X_LA-X_CB}{R_LA+R_CB}$ $\doteq 0$ when $X_L = X_C$ and R_L and R_C are small

FIG. 32b. Equivalent impedances of parallel combinations of L, C, and R.

Some commonly used parallel combinations of LCR are shown in Fig. 32a and b, with their impedance and phase relations.

13. Loss Due to Inserting Series or Shunt Impedance in Audio Circuits. In audio circuits, attenuation-frequency characteristics are often purposely modified by the insertion of corrective impedances such as equalizers, "tone controls," and scratch filters. The following formulas give the insertion losses in such cases:

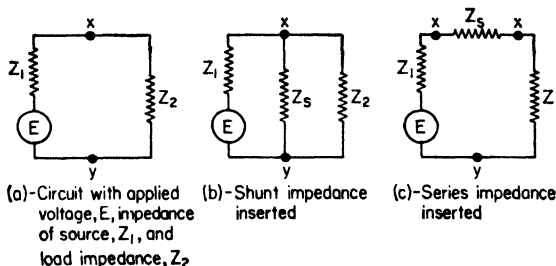



FIG. 33. Shunt and series impedances inserted in audio-frequency circuits.

1. *Shunt Impedance.* The loss due to inserting a shunt impedance Z (Fig. 33a and b) is

$$L = 20 \log_{10}\left(1 + \frac{Z_1Z_2}{Z_s(Z_1+Z_2)}\right) \quad \text{db} \tag{20}$$

The shunting impedance can usually be located at a point in the circuit where the imped-

ances Z_1 and Z_2 are matched and where each is substantially a pure resistance through the range of frequencies involved. Then, letting $Z_1 = Z_2 = R_0$, the loss is

$$L = 20 \log_{10} \left| \frac{2Z_s + R_0}{2Z_s} \right| \text{ db} = 20 \log_{10} \sqrt{1 + \frac{\cos \phi}{K} + \frac{1}{4K^2}} \text{ db} \qquad (20a)$$

where $K = |Z_s|/R_0$ and ϕ is the phase angle of Z_s. For various values of K and ϕ the loss can be read from the curve (Fig. 34).

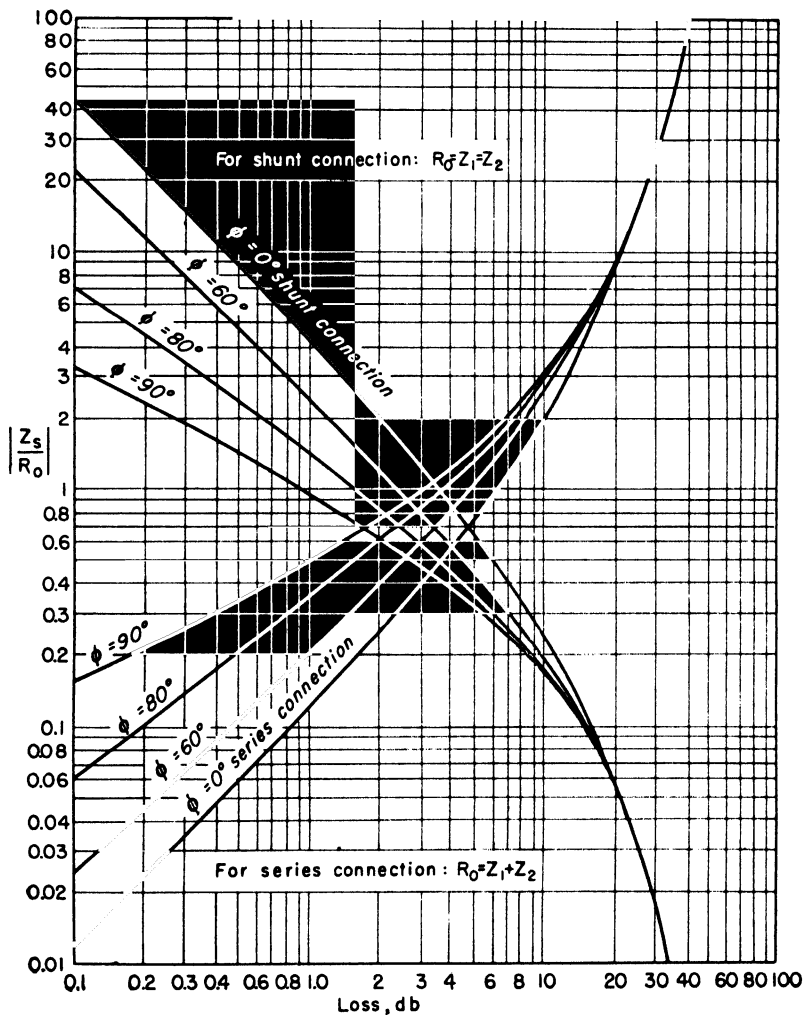



FIG. 34. Transmission loss due to insertion of shunt or series impedance.

2. *Series Impedance.* The loss in decibels due to inserting a series impedance Z_s (Fig. 33*a* and *c*) is

$$L = 20 \log_{10} \left(\frac{Z_1 + Z_2 + Z_s}{Z_1 + Z_2} \right) \text{ db} \qquad (21)$$

The series impedance can usually be inserted at a point in the circuit where the impedances Z_1 and Z_2 are matched and where each is substantially a pure resistance through the range of frequencies involved. Then, letting $Z_1 + Z_2 = R_0$, the loss is

$$L = 20 \log_{10} \left| \frac{R_0 + Z_s}{R_0} \right| \quad \text{db}$$

$$= 20 \log_{10} \sqrt{1 + 2K \cos \phi + K^2} \quad \text{db} \tag{21a}$$

where $K = |Z_s|/R_0$ and ϕ is the phase angle of Z_s. The loss can be read from Fig. 34 for various values of K and ϕ.

14. Series Resonance. With L and C in series, resonance is the condition where $X_L = X_C$ and current is maximum. In the ideal resistanceless circuit, the impedance would be zero at resonance; so series resonance is often referred to as a "zero" in network theory. The presence of resistance in an actual physical circuit does not affect the resonance frequency for a given L and C.

Resonance frequency:

$$f_r = \frac{1}{2\pi \sqrt{LC}} \tag{22}$$

where f is in cycles, L in henrys, and C in farads.

Impedance at resonance:

$$Z_r = R_r \quad \text{ohms} \tag{23}$$

Impedance at any frequency, f_1:

$$Z_1 = R_1 + j2\pi L \left(\frac{f_1^2 - f_r^2}{f_1} \right) = |Z_1| \underline{/\phi_1} \quad \text{ohms} \tag{24}$$

$$|Z_1| = \sqrt{R_1^2 + 4\pi^2 L^2 \left(\frac{f_1^2 - f_r^2}{f_1} \right)^2} \tag{25}$$

$$\phi_1 = \tan^{-1} \frac{X_1}{R_1} = \tan^{-1} \left[\frac{2\pi L(f_1^2 - f_r^2)}{f_1 R_1} \right] \tag{26}$$

Subscript$_{(r)}$ denotes values at resonance, and subscript$_{(1)}$ values at any frequency f_1.

Impedance in terms of Q:

Either of two conditions may be assumed: (1) Q is constant as frequency varies; (2) R is constant as frequency varies. Neither is exactly true in any practical case because R actually varies with the frequency in a complex manner. Expressions in terms of Q, however, give fair approximations if $Q = 10$ or more and are, therefore, useful.

The differences for impedance between constant R and constant Q are a few per cent in impedance near resonance, for Q on the order of 10. For large Q and frequencies farther from resonance, the impedances by either assumption tend to become equal. The differences in phase-angle results are largest at frequencies farthest from resonance and for small Q's. Near resonance and with large values of Q, the values by both assumptions approach equality.

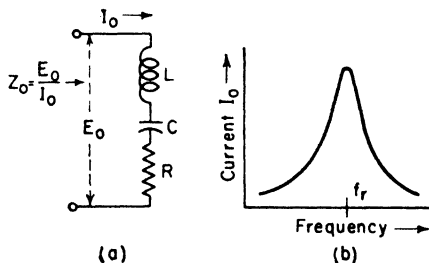



FIG. 35. (a) Series-resonant circuit and (b) current vs. frequency for constant E_0.

For *constant R:*

$$R = \frac{\omega_r L}{Q_r} = \frac{\omega_1 L}{Q_1} \tag{27}$$

$$\frac{|Z_1|}{\omega_r L} = \sqrt{\frac{1}{Q_r^2} + n^2 + \frac{1}{n^2} - 2} \tag{28}$$

$$\phi_1 = \tan^{-1}\left[Q_r\left(n - \frac{1}{n}\right)\right] \tag{29}$$

where $n = f_1/f_r$.

For *constant Q:*

$$Q = \frac{\omega_r L}{R_r} = \frac{\omega_1 L}{R_1}$$

$$\frac{|Z_1|}{\omega_r L} = \sqrt{\frac{n^2}{Q^2} + n^2 + \frac{1}{n^2} - 2} \tag{30}$$

$$\phi_1 = \tan^{-1}\frac{Q}{n}\left(n - \frac{1}{n}\right) \tag{31}$$

where $n = f_1/f_r$.

Subscript$_{(r)}$ denotes values at resonance, and subscript$_{(1)}$ values at any frequency f_1.

At resonance, the voltage drop (E_L) across L is numerically equal to the voltage drop (E_C) across C. In terms of applied voltage E_0 when $Q > 25$,

$$|E_L| = |E_C| = QE_0$$

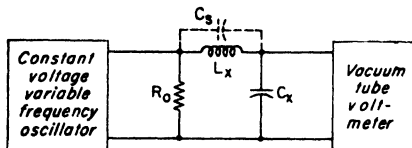
or

$$Q = \frac{|E_C|}{E_0} \tag{32}$$

Fig. 36. Direct-reading Q meter.

The latter equation is the basis of a direct-reading Q meter illustrated in Fig. 36. The inductance L_x to be measured is connected in series with a low-loss variable capacitor C_x. Voltage E_0 is applied from a constant-voltage variable-frequency oscillator terminated in a very small resistance R_0. The frequency and series capacitances are adjusted for series resonance. A high-input-impedance vacuum-tube voltmeter

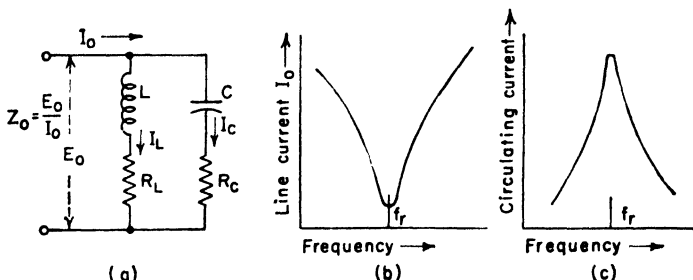



Fig. 37. Parallel-resonant circuit and characteristics for constant E_0 applied.

reads the series capacitor voltage $|E_C|$ and may be calibrated directly in terms of Q from Eq. (32), the input voltage being known and constant. In precision instruments, correction factors are applied to eliminate the loading effects of R_0 and the voltmeter impedance, and of the distributed capacitance C_s of the coil.

15. Parallel or Antiresonance. Resonance in parallel LC circuits, often called "antiresonance," or a "pole" of impedance, corresponds approximately to the con-

dition of maximum circulating current, or to maximum impedance between the terminals, as indicated in Fig. 37. Exact analysis, in each case based on slightly differing assumptions, yields various equations for the resonance frequency, all of which differ from each other by only small amounts. Those most useful in com-

Product LC for Any Frequency, F

(1)		(2)			(3)		(4)		(5)				(6)
F μμf-μh-kc μμf-h-cps		LC			F μf-h-cps μμf-h-kc		F μμf-μh-kc μμf-h-cps		LC				F μf-h-cps μμf-h-kc
100	0	253	30	30	10		400	0	15	83	14		40
105	0	229	75	30	10	5	410	0	15	06	85		41
110	0	209	34	10	11		420	0	14	35	96		42
115	0	191	53	30	11	5	430	0	13	69	95		43
120	0	175	90	40	12		440	0	13	08	38		44
125	0	162	11	20	12	5	450	0	12	50	88		45
130	0	149	88	30	13		460	0	11	97	08		46
135	0	138	98	60	13	5	470	0	11	46	67		47
140	0	129	23	60	14		480	0	10	99	40		48
145	0	120	47	70	14	5	490	0	10	55	00		49
150	0	112	58	00	15		500	0	10	13	21		50
155	0	105	43	20	15	5	520	0	9	36	77		52
160	0	98	94	60	16		540	0	8	68	67		54
165	0	93	04	05	16	5	550	0	8	37	37		55
170	0	87	64	80	17		560	0	8	07	73		56
175	0	82	71	10	17	5	580	0	7	52	98		58
180	0	78	18	00	18		600	0	7	03	62		60
185	0	74	01	10	18	5	620	0	6	58	95		62
190	0	70	16	70	19		640	0	6	18	42		64
195	0	66	61	48	19	5	650	0	5	99	53		65
200	0	63	32	57	20		660	0	5	81	50		66
210	0	57	43	83	21		680	0	5	47	80		68
220	0	52	33	53	22		700	0	5	16	94	5	70
230	0	47	88	34	23		720	0	4	88	62	5	72
240	0	43	97	62	24		740	0	4	62	57	0	74
250	0	40	52	84	25		750	0	4	50	31	6	75
260	0	37	47	09	26		760	0	4	38	54	4	76
270	0	34	74	66	27		780	0	4	16	34	3	78
280	0	32	30	90	28		800	0	3	95	78	6	80
290	0	30	11	93	29		820	0	3	76	71	5	82
300	0	28	14	48	30		840	0	3	58	98	9	84
310	0	26	35	83	31		850	0	3	50	59	2	85
320	0	24	73	66	32		860	0	3	42	48	6	86
330	0	23	26	01	33		880	0	3	27	09	6	88
340	0	21	91	20	34		900	0	3	12	72	0	90
350	0	20	67	77	35		920	0	2	99	27	1	92
360	0	19	54	50	36		940	0	2	86	67	2	94
370	0	18	50	28	37		950	0	2	80	66	8	95
380	0	17	54	18	38		960	0	2	74	85	1	96
390	0	16	65	37	39		980	0	2	63	74	7	98
400	0	15	83	14	40		1000	0	2	53	30	3	100

NOTES: The double rules in Cols. 1 and 4 indicate that the decimal point in Cols. 2 and 5 is at the double rule; the single rules in Cols. 3 and 6 indicate that the decimal point in Cols. 2 and 5 is at the first single rule. Thus, the LC product for 100 cycles (Col. 1) is 2,533,030 (Col. 2) when the units are μμf and h; or is 2.533030 (Col. 5) if the units are μf and h (Col. 6).

For frequencies not listed, linear interpolation may be used for accuracies of 0.1 %. For higher accuracies, $LC = 253303/f^2$, neglecting decimals.

munications work are listed below. In all cases, the units are ohms, henrys, farads, and cycles.

1. *Free Oscillation Resonance* (R_C negligible).

$$f_r = \frac{1}{2\pi\sqrt{LC}}\sqrt{1 - \frac{R_L^2C}{4L}} \tag{33}$$

This is the natural frequency of oscillation in a parallel circuit excited by a single pulse. (See Sec. 5 also.)

2. *L Variable and Adjusted for Maximum* $|Z_0|$* (R_C negligible). If R_L is constant as L varies,

$$f_r = \frac{1}{2\pi\sqrt{LC}}\sqrt{1 + \frac{R_L^2C}{L}} \tag{34}$$

If L/R_L is constant as L varies,

$$f_r = \frac{1}{2\pi\sqrt{LC}} \tag{35}$$

3. *Frequency Variable and Adjusted for Maximum* $|Z_0|$ (R_C negligible).

$$f_r = \frac{1}{2\pi\sqrt{LC}}\sqrt{\sqrt{1 + \frac{2R_L^2C}{L}} - \frac{R_L^2C}{L}} \tag{36}$$

4. *C Variable and Adjusted for Maximum* $|Z_0|$ (R_C negligible).

$$f_r = \frac{1}{2\pi\sqrt{LC}}\sqrt{1 - \frac{R_L^2C}{L}} \tag{37}$$

In this case, the phase angle of Z_0 is zero at resonance (Z_0 = pure resistance), and the line current I_0 is in phase with applied voltage E_0.

The differences between these resonance frequencies in ordinary circuits are of the order of a fraction of 1 per cent and are trivial except when accurate measurements or extremely critical circuit adjustments are involved. An example of the latter case is the adjustment of a transmitter tank circuit for the highest possible efficiency. Condition 4, which gives zero phase angle and unity power factor for the equivalent impedance of the circuit at resonance, is the most common definition of parallel resonance.

5. Z_0 = *Pure R, and R_C and R_L Not Negligible.*

$$f_r = \frac{1}{2\pi\sqrt{LC}}\sqrt{\frac{L - R_L^2C}{L - R_C^2C}} \tag{38}$$

6. Z_0 = *Pure R, and R_C Negligible.*

$$f_r = \frac{1}{2\pi\sqrt{LC}}\sqrt{1 - \frac{R_L^2C}{L}} = \frac{1}{2\pi\sqrt{LC}}\sqrt{1 - \frac{1}{Q_r^2}} \quad \text{(approx)} \tag{39}$$

where Q_r is the coil Q at resonance. This is the most useful expression for f_r for practical applications.

7. *Approximation When $R_L = R_C$ or When R_C and R_L Are Both Negligible.*

$$f_r = \frac{1}{2\pi\sqrt{LC}} \tag{40}$$

* $|Z_0|$ is the magnitude of the equivalent impedance.

This ignores circuit resistance entirely and is used when approximate results are satisfactory.

8. *Special Case When* $R_L = R_C = \sqrt{L/C}$. The circuit is resonant at all frequencies and $Z_0 = R_L = R_C = \sqrt{L/C}$ at all frequencies. Since L/C is usually of the order of 10^6, Z_0 will be about 1,000 ohms, which limits the practical applications of the circuit.

16. Impedance Relations in Parallel-resonant Circuits. Exact impedance and phase-angle equations are given in Fig. 32*b*. The following are practical approximations.

Impedance at resonance (R_C negligible):

$$Z_r = \frac{X_C^2}{R_L} = \frac{X_L^2}{R_L} = \frac{X_L X_C}{R_L} = \frac{L}{R_L C} \quad \text{ohms (resistive)} \tag{41}$$

Impedance at any frequency, f_1 (R_C negligible):

$$Z_1 = X_C \frac{R_L X_C - j(R_L^2 + X_L^2 - X_L C_C)}{R_L^2 + (X_L - X_C)^2} = |Z_1| \underline{/\phi_1} \tag{42}$$

$$|Z_1| = \frac{X_C \sqrt{R_L^2 + X_L^2}}{\sqrt{R_L^2 + (X_L - X_C)^2}} \tag{43}$$

$$\phi_1 = \tan^{-1}\left(-\frac{R_L^2 + X_L^2 - X_L X_C}{R_L X_C}\right) \tag{44}$$

If $R_L \ll X_L$,

$$|Z_1| = \frac{L}{C} \frac{1}{\sqrt{R_L^2 + (X_L - X_C)^2}} = \frac{L}{C} \frac{1}{\sqrt{R_L^2 + 4\pi^2 L^2 \left(\dfrac{f_1^2 - f_r^2}{f_1}\right)^2}} \tag{45}$$

Subscript$_{(r)}$ denotes values at resonance, and subscript$_{(1)}$ values at any frequency, f_1.

Impedance in terms of Q:

The comments in Sec. 14 relative to variation of R and Q in series-resonant circuits also apply here.

At resonance (R_C negligible, $Q > 25$):

$$Z_r = \frac{Q_r}{\omega_r C} = Q_r \omega_r L = R Q_r^2 \quad \text{ohms} \tag{46}$$

and

$$\frac{Z_r}{\omega_r L} = Q_r \tag{47}$$

where Q_r = coil Q at resonance.

At any frequency, f_1 (R_C negligible, $Q > 25$) for *constant R:*

$$\frac{|Z_1|}{\omega_r L} = \frac{\sqrt{\dfrac{1}{Q_r^2} + n^2}}{n\sqrt{\dfrac{1}{Q_r^2} + n^2 + \dfrac{1}{n^2} - 2}} \tag{48}$$

$$\phi_1 = \tan^{-1}[-nQ_r(n^2 - 1)] \tag{49}$$

where $n = f_1/f_r$.

Figure 38 is a family of resonance curves for $|Z_1|/\omega_r L$ vs. frequency ratio n, plotted for the condition of constant R. Figure 39 is a corresponding set of curves for ϕ_1.

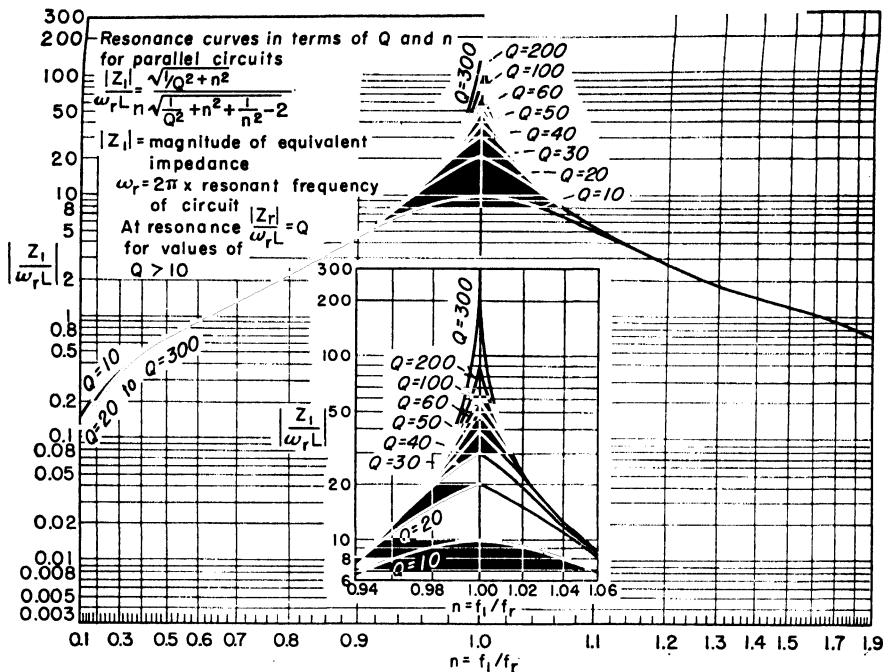



Fig. 38. Parallel-resonance curves.

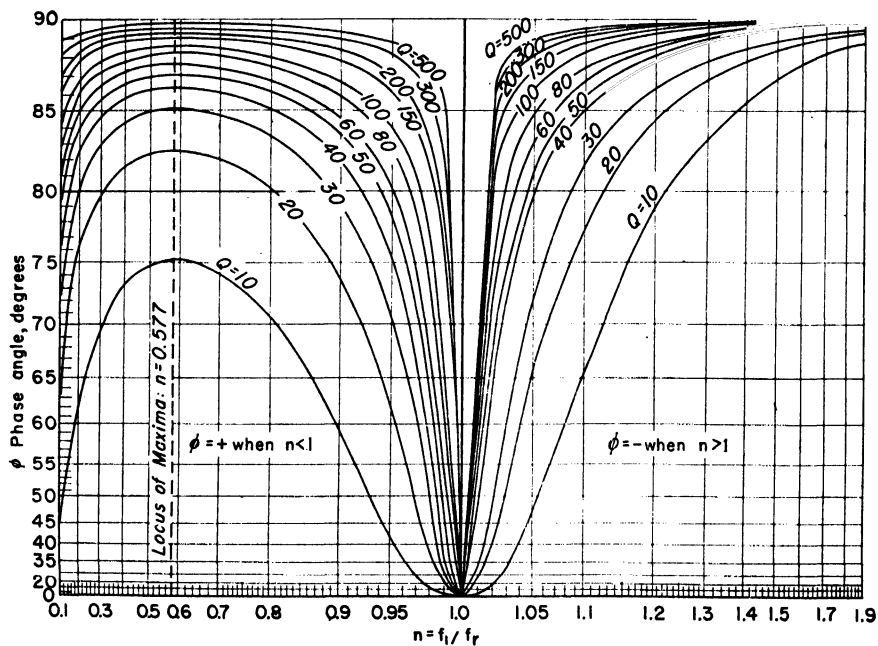



Fig. 39. Phase angle of parallel LC circuit in terms of n and Q.

For *constant* Q (R_C negligible, $Q > 25$):

$$\frac{|Z_1|}{\omega_r L} = \frac{\sqrt{(1/Q^2) + 1}}{\sqrt{(n^2/Q^2) + n^2 + (1/n^2) - 2}} \tag{50}$$

$$\phi_1 = \tan^{-1}[-Q(n^2 - 1)] \tag{51}$$

where $n = f_1/f_r$.

In all cases, ϕ_1 is positive when f_1 is less than f_r and is negative when f_1 is greater than f_r.

When the circuit Q is 25 or larger, the impedance of a parallel LCR circuit at $f_1 = 0.707 f_r$ and at $f_1 = 1.414 f_r$ is

$$|Z_1| = \frac{1.414}{Q_r}|Z_r| \qquad \text{or} \qquad \frac{|Z_1|}{|Z_r|} = \frac{1.414}{Q_r} \tag{52}$$

This is the basis for one method of measuring the Q of an inductance coil. The coil is connected in parallel with a low-loss variable capacitor, and voltage is applied from a constant-voltage variable-frequency source. The line current (I_r) is measured at resonance frequency and at 70.7 per cent or 141.4 per cent of resonance frequency (I_1), and Q is then determined from the relation

$$Q = 1.414\frac{|I_1|}{|I_r|}$$

17. Line and Circulating Currents in Parallel-resonant Circuits. The true circulating current in a parallel-resonant circuit operating at or very near resonance frequency is the current representing the energy oscillating to and fro between L and C. This is a relatively large current (as compared to line current) and is a component of both the L branch and the C branch currents I_L and I_C (see Fig. 37). A second component of I_L is the current representing the losses in the L branch, which include the I^2R_L loss due to the equivalent ohmic resistance of the branch plus the energy taken from the circuit by any inductively coupled load. A second component of I_C represents ohmic losses in the C branch plus any load that may be coupled to that branch.

In addition, if the applied voltage E_0 is not sinusoidal and of frequency $= f_r$, the harmonic currents through the L and C branches will not be equal because the reactances of the two branches are equal (approximately) only at the resonance frequency f_r. Thus currents produced by components of E_0 whose frequency is less than f_r will flow principally through the L branch, while those whose frequency is greater than f_r will flow principally through the C branch. Hence the total I_L is not actually equal to the total I_C and the difference is the line current, I_0.

In some cases, when the total losses of the LCR circuit are small, it is assumed that the circulating current is $I_L = I_C$. At the same time the line current cannot be ignored but is given by the approximation

$$I_0 = \frac{I_L}{Q} = \frac{I_C}{Q} \tag{53}$$

or, in terms of I_0,

$$I_L = I_C = QI_0 \tag{53a}$$

This approximation is most nearly accurate when the applied voltage is sinusoidal and has frequency f_r and when the circuit Q is high; it is inaccurate under any other conditions.

18. Design of Parallel-resonant Circuits. To design a parallel-resonant circuit, we have to determine values of L, C, R_L, and R_C to satisfy a given set of conditions. Values of $|Z_r|$ at resonance, the resonance frequency, and Q will first have to be determined by analysis of the intended use of the resonant circuit.

In a vacuum-tube oscillator, for example, f_r of the tuned circuit is known, and $|Z_r|$ at resonance is fixed by the permissible plate voltage swing. For Q (which includes the effect of the external load coupled to the tuned circuit, as well as the latter's ohmic resistance) a value of from 12 to 20 represents a good compromise between oscillator efficiency and frequency stability.

Another example of the factors involved in the choice of Q in an application is that of a tuned circuit for an r-f amplifier to pass a modulated carrier. In this case the LC circuit must have sufficient decrement to damp out its own natural oscillations between successive peaks of modulation; otherwise there is an effective decrease in modulation percentage with a corresponding loss of fidelity. If the carrier frequency is f_C and the modulation frequency f_m, the maximum decrement of the modulated carrier wave at 100 per cent modulation is approximately

$$\delta_1 = 2.303 \log_{10}\left(\frac{1}{1 - \pi\frac{f_m}{f_C}}\right) \tag{54}$$

The decrement δ_2 of the tuned circuit should be 10 to 20 times as large as δ_1 for faithful response. Then Q for the tuned circuit is

$$Q = \frac{\pi}{\delta_2} \tag{55}$$

The value of $|Z_r|$ at resonance will depend upon plate-load impedance requirements of the amplifier tube.

In some cases the ratio of volt-amperes circulating in LC to watts dissipated is the basis for the design of an LC circuit; in this case

$$Q = \frac{\text{volt-amp stored}}{\text{watts dissipated}} \tag{56}$$

The effect of any load coupled to a tuned circuit must be taken into account as part of the total effective R of the circuit. If the power taken by the load is W_d watts and I_C is the circulating current in LCR, the total equivalent impedance of the circuit is, approximately,

$$R = R_0 + \frac{W_d}{I_C^2} \quad \text{ohms} \qquad \text{(approx)} \tag{57}$$

where R_0 is the ohmic resistance.

Examples of Design of Parallel-resonant Circuit. Assume that a parallel circuit (Fig. 37) is to be resonant at 5,000 cycles, with an impedance of 4,000 ohms at resonance ($n = 1$) and an impedance of 100 ohms at 3,000 cycles ($n = 0.6$). From Fig. 38, $|Z_1|/\omega_r L = 0.9$ for all values of Q when $n = 0.6$. At resonance $|Z_r|/\omega_r L$ is to be $(4{,}000/100) \times 0.9 = 36$. From the curves it is found that $Q = 36$ gives $|Z_r|/\omega_r L = 36$ at $n = 1$ where $\omega_r = 31{,}416$. Then for $n = 1$,

$$Z_r = 36\omega_r L = 4{,}000 \quad \text{or} \quad L = \frac{4{,}000}{36 \times 31{,}416} = 0.00354 \quad \text{henry}$$

LC for 5,000 cycles $= 10.136 \times 10^{-10}$. Then $C = LC/L = 0.286 \times 10^{-6}$ farad, and $R = \omega_r L/Q = 3.08$ ohms.

As a second example suppose there is to be designed a tuned circuit for an r-f amplifier which requires a plate-load impedance of 10,000 ohms and which is to amplify a 1,000-kc carrier with amplitude modulation up to 5,000 cycles. From Eq. (54),

$$\delta_1 = 2.303 \log_{10} \frac{1}{1 - \pi\frac{5 \times 10^3}{10^6}} = 0.0159$$

$$\delta_2 = 20\,\delta_1 = 0.318$$

$$Q = \frac{\pi}{\delta_2} = 9.85 \doteq \frac{|Z_r|}{\omega_r L} \qquad \text{from Eq. (55)}$$

$$\omega_r L = \frac{|Z_r|}{Q} = \frac{10{,}000}{9.85} = 1{,}015$$

$$L = \frac{\omega_r L}{\omega_r} = \frac{1{,}015}{2\pi f_c} = 162 \qquad \mu\text{h}$$

$$LC \text{ for } 1{,}000 \text{ kc} = 2.53 \times 10^{-10}$$

$$C = \frac{LC}{L} = \frac{2.53 \times 10^{-10}}{162 \times 10^{-6}} = 157 \qquad \mu\mu\text{f}$$

$$R = \frac{\omega_r L}{Q} = \frac{1{,}015}{9.85} = 103 \qquad \text{ohms}$$

This consists of the ohmic resistance of LCR plus the equivalent R of the coupled load, as computed by Eq. (57).

19. Tapped Tank Circuits. In some cases the high impedance of a parallel LCR circuit at resonance is a disadvantage, *e.g.*, at the end of a low-impedance transmission line where the correct termination is about 500 ohms. However, the low impedance can be obtained by tapping the LCR circuit in either the L or C branch as shown in Fig. 40. The result is a coupled circuit, that part of the reactance between B and C being the mutual impedance.

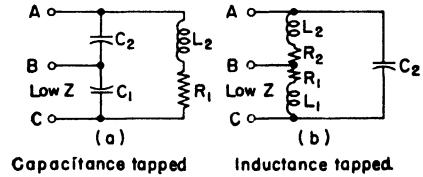



FIG. 40. Tapped tank circuits.

1. *Capacitance Tapped.* In Fig. 40a, the impedance at B-C is

$$|Z_{BC}| = \frac{\sqrt{L^2C_2^2\left(\frac{1}{C_1(C_1 + C_2)}\right)^2 + \frac{R^2LC_2}{C_1(C_1 + C_2)}}}{R} \tag{58}$$

If R is small,

$$|Z_{BC}| = \frac{LC_2}{RC_1(C_1 + C_2)} \tag{58a}$$

and its ratio to the impedance Z_{AC} is

$$\frac{|Z_{BC}|}{|Z_{AC}|} = \frac{C_2^2}{(C_1 + C_2)^2} \tag{59}$$

The resonant frequency is

$$f_r = \frac{1}{2\pi\sqrt{L\frac{C_1C_2}{C_1 + C_2}}} \tag{60}$$

and the impedances Z_{AC} and Z_{BC} are both purely resistive at resonance.

The ratio of C_1 to C_2 for a given ratio between Z_{AC} and Z_{BC} is

$$\frac{C_1}{C_2} = \left(\sqrt{\frac{Z_{AC}}{Z_{BC}}} - 1\right) \tag{61}$$

In terms of the resonant frequency, inductance, and the impedance ratio,

$$C_1 = \frac{1}{4\pi^2 f_r^2 L}\sqrt{\frac{Z_{AC}}{Z_{BC}}} \tag{62}$$

$$C_2 = \frac{1}{4\pi^2 f_r^2 L\left(1 - \sqrt{\frac{Z_{BC}}{Z_{AC}}}\right)} \tag{63}$$

2. *Inductance Tapped.* In Fig. 40*b* the inductance is tapped, and the impedance at *B-C* is (assuming no mutual inductance between L_1 and L_2)

$$|Z_{BC}| = \frac{\sqrt{\left(R_1R_2 - \frac{L_1L_2}{(L_1 - L_2)C_2} + \frac{L_2}{C_2}\right)^2 + \left(\frac{R_2L_1}{\sqrt{(L_1 + L_2)C_2}} + \frac{R_1L_2}{\sqrt{(L_1 + L_2)C_2}} - \frac{R_1\sqrt{(L_1 + L_2)C_2}}{C_2}\right)^2}}{R_1 + R_2} \tag{64}$$

If R_1 and R_2 are small,

$$|Z_{BC}| = \frac{L_2^2}{C_2(R_1 + R_2)(L_1 + L_2)} \tag{64a}$$

and its ratio to the total impedance Z_{AC} is

$$\frac{|Z_{BC}|}{|Z_{AC}|} = \frac{L_2^2}{(L_1 + L_2)^2} \tag{65}$$

The resonant frequency is

$$f_r = \frac{1}{2\pi\sqrt{(L_1 + L_2)C_2}} \tag{66}$$

and the impedances Z_{AC} and Z_{BC} are both resistive at resonance.

The ratio of L_1 to L_2 for a given ratio between Z_{AC} and Z_{BC} is

$$\frac{L_1}{L_2} = \sqrt{\frac{Z_{AC}}{Z_{BC}}} - 1 \tag{67}$$

In terms of the frequency, capacitance, and the impedance ratio,

$$L_1 = \frac{1}{4\pi^2 f_r^2 C_2}\left(1 - \sqrt{\frac{Z_{BC}}{Z_{AC}}}\right) \tag{68}$$

$$L_2 = \frac{1}{4\pi^2 f_r^2 C_2}\sqrt{\frac{Z_{BC}}{Z_{AC}}} \tag{69}$$

20. Resonant Circuits in Tandem. In a network of pure reactances, the number of poles is never different by more than one from the number of zeros (including the zero and pole at each extreme, one at zero and the other at infinite frequency), and the zeros or poles occur alternately with respect to increasing frequency.

Thus when two parallel-resonant circuits are connected in series as in Fig. 41*a*, a new zero, or series resonance, is introduced at some frequency between the two original poles. The new resonance is at

$$f = \frac{1}{2\pi}\sqrt{\frac{L_1 + L_2}{L_1L_2(C_1 + C_2)}} \tag{70}$$

Similarly, with two series-resonant circuits in parallel as in Fig. 41*b*, a pole appears between their zeros at

$$f = \frac{1}{2\pi}\sqrt{\frac{C_1 + C_2}{C_1C_2(L_1 + L_2)}} \tag{70a}$$

Other cases of multiple resonance involving resonant circuits and reactances in tandem are listed among the inverse two-terminal networks of Sec. 36.

Multiple resonance imposes practical limits on the way in which resonant circuits may be combined, since unwanted resonances may occur unless the branches are isolated from each other. It prevents, for example, the use of more than one parallel LC peaking circuit in the plate circuit of a single amplifier stage. In this case, the necessary isolation may be achieved by placing the other peaking circuits in different stages. Inserting resistance between resonant branches often solves the isolation problem in passive networks, but at the cost of additional losses.

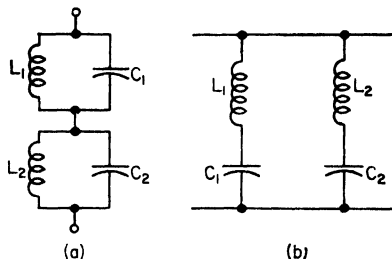



FIG. 41. Resonant circuits in tandem.

There are, of course, many applications where multiple resonance serves a useful purpose, notably in the design of impedance branches for four-terminal wave filters and corrective networks.

COUPLED CIRCUITS

21. Coupling. Two circuits are coupled if they have a common impedance so situated that it causes the current flowing in one circuit to influence the current in the other. The common impedance may be resistance, capacitance, inductance, mutual inductance, or any combination of these. In the general case where it contains both resistance and reactance, the analysis is complicated.[1]

22. Coefficient of Coupling. The extent of coupling between two circuits is called the *coefficient of coupling* and is

$$K = \frac{X_m}{\sqrt{X_1X_2}} \tag{71}$$

where X_m is any one component of the mutual impedance (ωL_m, $1/\omega C_m$, or ωM in ohms) and X_1 and X_2 are the total impedance components of the same kind in the respective circuits, including the mutual impedance. K varies in value between 0 and 1; if it is nearly 1, the coupling is *close* or *tight;* if it is near 0 ($K = 0.1$ or less), the coupling is *loose.*

23. Coupling Coefficients for R-f Circuits. Values of K for four types of coupling most used in r-f circuits are given in Fig. 42.

24. Mutual Inductance Coupling at Low Frequencies. In Fig. 43 circuit I is coupled to circuit II by the transformer (mutual inductance) T. The primary and secondary windings, X_p and X_s, are assembled on a closed iron core. At low frequencies, including audio and commercial power frequencies, the magnetic coupling is almost perfect, which results in a large value of M and makes K nearly equal to 1. Also X_p and X_s have sufficient inductance so that $X_p \gg Z_1$ and $X_s \gg Z_2$.

Under these conditions, the following approximate relations hold:

Impedance looking into primary:

$$Z_0 = \frac{X_p}{X_s} Z_2 = \left(\frac{N_p}{N_s}\right)^2 Z_2$$

[1] For discussions of this case, see G. W. Pierce, "Electric Oscillations and Electric Waves," McGraw-Hill, 1920, and C. B. Aiken, Two-mesh Tuned Coupled Filters, *Proc. IRE*, February, 1937, p. 230.

where N_p and N_s are number of turns in primary and secondary windings, respectively.

Current ratio:

$$\frac{I_1}{I_2} = \frac{N_s}{N_p} \qquad I_2 = \frac{N_p}{N_s} I_1 \qquad I_1 = \frac{N_s}{N_p} I_2$$

Voltage ratio:

$$\frac{E_1}{E_2} = \frac{N_p}{N_s} \qquad E_1 = \frac{N_p}{N_s} E_2 \qquad E_2 = \frac{N_s}{N_p} E_1$$

Volt-amperes:

$$E_1 I_1 = E_2 I_2 + \text{transformer losses}$$

Impedance ratio:

$$\frac{Z_0}{Z_2} = \left(\frac{N_p}{N_s}\right)^2 \qquad \frac{Z_2}{Z_0} = \left(\frac{N_s}{N_p}\right)^2$$

These simple relations are not quite exact for any actual transformer because of core losses, flux leakage, and resistance and distributed capacitance of windings. Also,

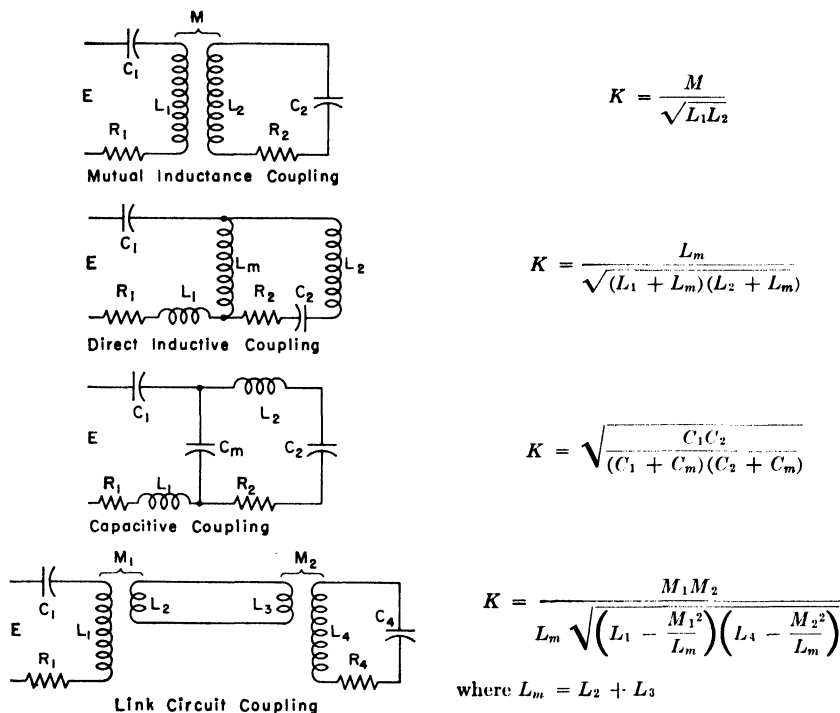



FIG. 42. Commonly used types of coupled circuits and their coupling coefficients.

since these imperfections vary with frequency, the various ratios likewise change with frequency. A well-designed transformer may give results approaching 90 to 95 per cent of the values given.

To minimize the effects of these imperfections, transformer designs must take into account not only the ratios but also the actual values of primary and secondary currents and impedances. For this reason, a-f transformers are rated in terms of the actual impedances between which they are designed to operate, and the frequency

range over which they will give satisfactory performance. Their performance outside these ratings is likely to be deficient. Thus, an a-f transformer designed to match one impedance of 500 ohms to another of 50 ohms (Z ratio of 10:1) probably would not give satisfactory results if used between impedances of 50,000 and 5,000 ohms, even though the impedance ratio is the same.

25. Mutual Inductance Coupling at Radio Frequencies. Excessive core losses, high resistance of windings, and large distributed capacitance bar the use of a-f types of iron-cored transformers in r-f circuits. Without iron cores, only small values of coupling between primary and secondary can be realized. Open cores of powdered iron are frequently used in r-f transformers, but their effect is principally to increase the circuit Q with little effect on the coupling. However, the range of frequencies that an r-f transformer is required to pass efficiently is usually small in comparison with the center frequency. The side bands of an a-f modulated signal, for instance, occupy a frequency band whose width is only 2 or 3 per cent of the carrier frequency. (This is not true for all types of radio signals, however. A television signal may require a band width of 10 to 15 per cent of the carrier frequency. In this case, simple transformer circuits are not suitable coupling devices.)

FIG. 43. Transformer coupling of low-frequency circuits.

When the band width is a small percentage of the center frequency, the limited values of K in an r-f transformer can be compensated by resonating the primary and secondary circuits. The transfer of energy in such an arrangement will vary with frequency, but this can be made reasonably constant over a narrow band that is still sufficiently wide to accommodate side bands.

26. Practical Relations in Transformer-coupled R-f Circuits. A practical application of a tuned coupled r-f circuit is shown in Fig. 44, which represents an i-f transformer used to couple the output of a pentode tube to the grid of the following tube.

FIG. 44. R-f amplifier stage with tuned coupled circuits.

The same circuit and the formulas pertaining to it apply equally well to t-r-f coils with untuned primaries if C_p is eliminated from both circuit and equations.

For practical purposes, the effectiveness of the coupled circuit of Fig. 44 is determined by its gain and its selectivity. These quantities are dependent not only upon the L, R, and C of the coupling circuits themselves, but also upon the effective input and output impedances of the two tubes, which shunt the primary and secondary and alter their resistance and capacitance.

The circuit of Fig. 44 is very nearly equivalent to that of Fig. 45 if

$$C_1 = C_p + C_o \qquad L_1 = L_p \qquad R_1 = \frac{1}{\omega^2 C_1^2 r_p} + r_p$$

$$C_2 = C_s + C_i \qquad L_2 = L_s \qquad R_2 = \frac{1}{\omega^2 C_2^2 R_i} + R_i$$

where C_o = output capacitance of tubes
C_i = input capacitance of tubes
r_p = tube plate resistance
R_i = tube input ohmic resistance
L_1 and L_2 = self-inductances of primary and secondary coils, each determined with the opposite circuit removed or uncoupled

For pentode tubes, the circuit input voltage is approximately

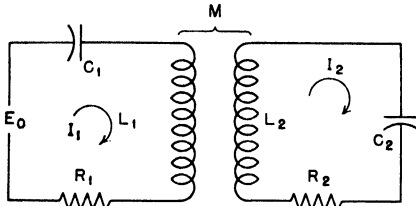



FIG. 45. Coupling-circuit equivalent of Fig. 44.

$$E_0 = \frac{G_m}{\omega C_2} \times E_i \qquad (72)$$

where E_i = input voltage to grid of first tube. The equivalent voltage E_0 may be considered as coming from a source having zero internal impedance.

For triodes, the equivalent voltage is approximately

$$E_0 = \mu \times E_i \qquad (73)$$

but in this case it must be considered as coming from a source whose internal impedance is equal to r_p, the plate resistance of the first tube. E_g is the actual input voltage to the grid of the second tube.

27. Band-pass Characteristics of Coupled Circuits. If the primary and secondary circuits are individually tuned to the same frequency f_0 so that $L_1C_1 = L_2C_2$ (L_1 not necessarily equal to L_2, nor C_1 to C_2), the secondary current I_2 varies with frequency according to one of the resonance curves of Fig. 46. Since E_g applied to the grid of the following tube is nearly proportional to I_2 for narrow bands, these curves also represent output voltage and, therefore, the gain vs. frequency characteristic of the amplifier stage.

FIG. 46. Variation of secondary current and gain with frequency and with degree of coupling.

The *shape* of the response curve depends upon the ratio of X_m^2 to $(R_1^2 + R_2^2)/2$,* while the *peak gain* depends upon the ratio of X_m^2 to R_1R_2. With mutual inductance coupling, $X_m = \omega M$.

1. *Peak Values of Secondary Current and Voltage vs. Coupling.*

CRITICAL COUPLING: When $X_m^2 = R_1R_2$, the coupling is called *critical*, and the curve has its maximum at f_0, as shown by curve 2, Fig. 46. Critical coupling also gives the maximum possible peak value, and this is

$$I_{2\,\max} = \frac{E_0}{2\sqrt{R_1R_2}} \quad \text{(approx)} \qquad (74)$$

or

$$E_{g\,\max} = \frac{E_0}{2\omega_0C_2\sqrt{R_1R_2}} \quad \text{(approx)} \qquad (74a)$$

where $\omega_0 = 2\pi f_0 = 1/\sqrt{L_1C_1} = 1/\sqrt{L_2C_2}$, and other quantities refer to the equivalent circuit of Fig. 45.

* This was pointed out by Aiken, *loc. cit.*

In terms of K and Q, the coupling is critical when $K = 1/\sqrt{Q_1Q_2}$, where Q_1 and Q_2 are the respective Q's of the primary and secondary circuits. Practically, these are approximately the Q's of the primary and secondary coils. If $Q_1 = Q_2 = Q, K = 1/Q$ for critical coupling.

UNDERCOUPLING: When $X_m^2 < R_1R_2$, the circuits are *undercoupled.* In this case, the curve has a lower maximum than for the case of critical coupling, as shown by curve 1, Fig. 46.

OVERCOUPLING: When $X_m^2 > R_1R_2$, the circuits are *overcoupled.* This may produce either a single or double peak, the latter case being shown by curve 3. In either case, the peak or peaks are equal to the maximum for critical coupling if $R_1 = R_2$, and slightly less if $R_1 \neq R_2$.

2. *Frequency Characteristic and Shape of Curves.*

SINGLE PEAK CURVE: When $X_m^2 \leqq (R_1^2 + R_2^2)/2$, the curve has a single peak (curves 1 and 2, Fig. 46) and has approximately the same shape as the resonance curve of a parallel LCR circuit whose $Q = \sqrt{Q_1Q_2}$.

DOUBLE PEAK CURVE: When $X_m^2 > (R_1^2 + R_2^2)/2$, the curve has a double peak (curve 3). If, in addition, $R_1 = R_2 = R$, the value of coupling for transition from the condition of single peak to double peak is the same as the value for critical coupling, and $X_m = R$. Furthermore, the peaks have maximum possible value.

3. *Selectivity and Gain.* Increasing coupling increases gain, up to the point of critical coupling, and reduces selectivity and, in practice, the best value of X_m is a compromise beween gain and selectivity. For example, in i-f transformers for a-f modulated signals, it is common practice to adjust coupling to about 80 per cent of critical.

In i-f circuits where high fidelity of response is of major importance, overcoupling is frequently used to broaden the nose of the curve by double peaking. This passes efficiently the upper side bands due to higher a-f modulating frequencies. Such a curve is shown in curve 3 of Fig. 46 and is reproduced in Fig. 47 with dimensions added for reference in Sec. 28.

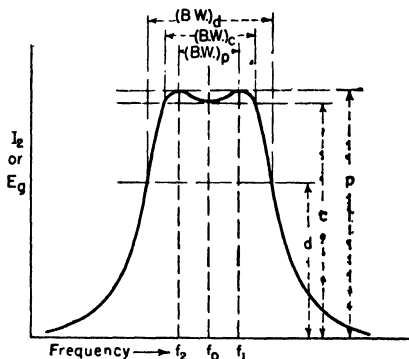



FIG. 47. Band-pass characteristic of overcoupled circuit.

28. Equations for Band Width and Gain. Although it is ordinarily easier to measure the response curve of a coupled circuit than it is to measure R_1 and R_2 and calculate the curve, relations between circuit parameters and gain and response are useful for design and analysis. The following relations give close approximations to exact values for the circuits of Figs. 44 and 45 and are based partly on Aiken's[1] results. The principal assumptions are that the band width is negligible compared to f_0, that R_1 and R_2 are constant throughout the band width, and that E_g is proportional to I_2.

Symbols used in gain and band-width equations:

E_i = input voltage to first tube grid (Fig. 44)

E_g = input voltage to second tube grid

$$\omega_0 = \frac{1}{\sqrt{L_1C_1}} = \frac{1}{\sqrt{L_2C_2}}, \text{ both circuits tuned to } f_0 = \frac{\omega_0}{2\pi}$$

$$K = \frac{M}{\sqrt{L_1L_2}}$$

[1] *Ibid.*

$$Q_1 = \frac{\omega_0 L_1}{R_1}$$

$$Q_2 = \frac{\omega_0 L_2}{R_2}$$

G_m = tube transconductance

Gain at center frequency for any value of coupling is proportional to dimension c at f_0 of Fig. 47 for either single or double peak curves and is

$$\frac{E_g}{E_i} = \frac{G_m M}{\omega_0 C_1 C_2 (R_1 R_2 + \omega_0^2 M^2)} = \frac{G_m \omega_0 K \sqrt{L_1 L_2}}{\dfrac{1}{Q_1 Q_2} + K^2} \tag{75}$$

If both circuits are identical and $R_1 = R_2 = R$, $Q_1 = Q_2 = Q$, and $L_1 = L_2 = L$, the gain is

$$\frac{E_g}{E_i} = \frac{G_m M}{\omega_0 C_1 C_2 (R^2 + \omega_0^2 M^2)} = \frac{G_m \omega_0 K L}{\dfrac{1}{Q_2} + K^2} \tag{76}$$

Maximum possible gain at center frequency is proportional to dimension p at f_0 and occurs when coupling is critical ($\omega_0 M = R_1 R_2$ or $K = 1/\sqrt{Q_1 Q_2}$) for either single or double peak curves and is

$$\frac{E_g}{E_i}\max = \frac{G_m \omega_0^2 L_1 L_2}{2\sqrt{R_1 R_2}} = \frac{G_m Q_1 Q_2 \sqrt{R_1 R_2}}{2} \tag{77}$$

If both circuits are identical and $R_1 = R_2 = R$, $Q_1 = Q_2 = Q$, and $L_1 = L_2 = L$, the gain is

$$\frac{E_g}{E_i}\max = \frac{G_m \omega_0^2 L^2}{2R} = \frac{G_m Q^2 R}{2} = \frac{G_m Q \omega_0 L}{2} \tag{78}$$

Gain at peaks of double peak curve (when $2\omega_0^2 M^2 > R_1^2 + R_2^2$) is the dimension p at f_1 and f_2 of Fig. 47 and is

$$\frac{E_g}{E_i} = \frac{2G_m M}{\omega_0 C_1 C_2 \sqrt{4\omega_0^2 M^2 (R_1 + R_2)^2 - (R_1^2 - R_2^2)}} \tag{79}$$

If $R_1 = R_2 = R$, the gain is

$$\frac{E_g}{E_i} = \frac{G_m}{2\omega_0^2 C_1 C_2 R} = \frac{Q_1 Q_2 R G_m}{2} = \frac{Q_1 \omega_0 L_1 G_m}{2} \tag{80}$$

Transition from single to double peak occurs when

$$f_0 = \frac{1}{2\pi M}\sqrt{\frac{R_1^2 + R_2^2}{2}} \quad \text{or} \quad M = \frac{1}{\omega_0}\sqrt{\frac{R_1^2 + R_2^2}{2}} \tag{81}$$

If $R_1 = R_2 = R$, transition occurs when

$$f_0 = \frac{R}{2\pi M} \quad \text{or} \quad M = \frac{R}{\omega_0} \tag{82}$$

Band width, peak to peak, **when $L_1 = L_2 = L$ is the dimension $(BW)_p$ in Fig. 47 and is**

$$(BW)_p = \frac{0.1592}{L}\sqrt{\omega_0^2M^2 - \frac{R_1^2 + R_2^2}{2}} = f_0\sqrt{K^2 - \frac{1}{2}\left(\frac{1}{Q_1^2} + \frac{1}{Q_2^2}\right)} \tag{83}$$

If $R_1 = R_2 = R$,

$$(BW)_p = \frac{0.1592}{L}\sqrt{\omega_0^2M^2 - R^2} = f_0\sqrt{K^2 - \frac{1}{4Q^2}} \tag{84}$$

If Q_1 and Q_2 are both large,

$$(BW)_p = Kf_0 \qquad \text{(approx)} \tag{85}$$

Band width between points on curve where gain equals gain at center frequency is $(BW)_c$ in Fig. 47 and is

$$(BW)_c = \sqrt{2}\,(BW)_p = \frac{0.225}{L}\sqrt{\omega_0^2M^2 - \frac{R_1^2 + R_2^2}{2}} \tag{86}$$

$$= \sqrt{2}f_0\sqrt{K^2 - \frac{1}{2}\left(\frac{1}{Q_1^2} + \frac{1}{Q_2^2}\right)} \tag{87}$$

If $R_1 = R_2 = R$,

$$(BW)_c = \frac{0.225}{L}\sqrt{\omega_0^2M^2 - R^2} = \sqrt{2}f_0\sqrt{K^2 - \frac{1}{Q_2}} \tag{88}$$

If Q_1 and Q_2 are both large,

$$(BW)_c = \sqrt{2}\,Kf_0 \qquad \text{(approx)} \tag{89}$$

Band width at point where gain is n db down from center gain of double peak curve when $L_1 = L_2 = L$ is $(BW)_d$ in Fig. 47:

$$(BW)_d = \frac{0.1125\sqrt{2\omega_0^2M^2 - R_1^2 - R_2^2 + \sqrt{(2\omega_0^2M^2 - R_1^2 - R_2^2)^2 + 4D(R_1R_2 + \omega_0^2M^2)^2}}}{L} \tag{90}$$

where D is a function of n to be read from Fig. 48. Equation (90) is approximate and is not accurate for points more than about 10 db down from center gain.

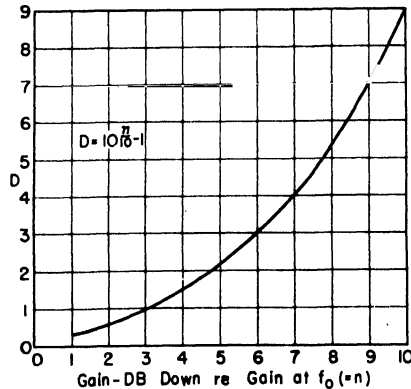



FIG. 48. Values of D vs. n.

If $R_1 = R_2 = R$,

$$(\text{BW})_d = \frac{0.1592\sqrt{\omega_0^2M^2 - R^2 + \sqrt{(\omega_0^2M^2 - R^2)^2 + D(R^2 + \omega_0^2M^2)^2}}}{L} \tag{91}$$

Coupling for band width of W cycles at n db down from center gain of double peak curve when $L_1 = L_2 = L$ is

$$\omega_0 M = 2.507\sqrt{\frac{LW}{D}\sqrt{39.48L^2W^2(2D+1) + D(R_1+R_2)^2} - \frac{39.48L^2W^2}{D} - R_1R_2} \tag{92}$$

where D is a function of n to be read from Fig. 48. W is identical with $(\text{BW})_d$ in other equations. Equation (92) is approximate and is not accurate for n larger than about 10.

SOME SPECIAL APPLICATIONS OF *LCR* CIRCUITS

29. Band-pass R-f Circuits. If two identical tuned circuits are capacitively or inductively coupled (Fig. 49), the circuit acts as a band-pass filter with a band width approximately

$$f_s = f_1 - f_2 = \frac{\sqrt{X_m^2 - R^2}}{2\pi L} \tag{93}$$

The band width varies with the tuning, increasing with the frequency in the inductive case, and decreasing with the frequency in the capacitive case (Fig. 49). These

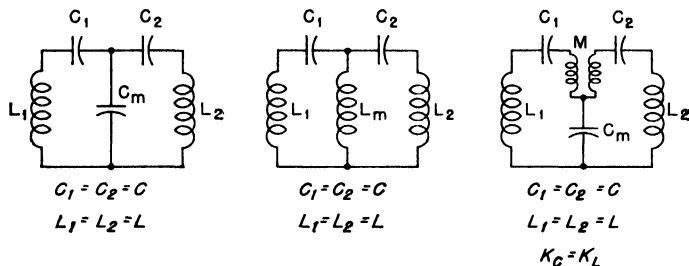



FIG. 49. Coupled circuits as band-pass filters.

opposing effects may be combined in the manner shown in Fig. 49, so that the band width is maintained substantially constant while the circuits are tuned over a wide range of frequency by adjustment of C_1 and C_2.

Uehling[1] has shown that this condition obtains when

$$X_{m_n} = \pm\sqrt{R_n^2 + 4\pi^2L^2f_s^2} \tag{94}$$

where R_n is the resistance and L the total inductance of each branch and f_s is the band width. With X_m computed for the two boundary frequencies f_a and f_b of the tuning range, the values of M and C_m required are given by

$$M = \frac{X_{m_b}f_b - X_{m_a}f_a}{2\pi(f_a^2 - f_b^2)} \tag{95}$$

$$C_m = \frac{f_a^2 - f_b^2}{2\pi f_a f_b(X_{m_a}f_b - X_{m_b}f_a)} \tag{96}$$

[1] *Electronics*, September, 1930, p. 279.

Representative values of M and C_m for $f_a = 1{,}500$ kc, $f_b = 550$ kc, $R_a = 30$ ohms, $R_b = 10$ ohms, $L = 200 \times 10^{-6}$ henry, and $f_s = 10$ kc, which are typical constants of broadcast circuits, are

$$M = 3.2 \times 10^{-6} \quad \text{henry} \quad \text{and} \quad C_m = 0.06 \quad \mu\text{f}$$

The inductive coupling M must be negative so that its effect will be additive to that of C_m. This may be obtained by winding the coils M (Fig. 49) of two wires side by side and by connecting the "start" ends of the coils to C_1 and C_2 and the "finish" ends to C_m.

30. Decoupling Filters. When the plate current for several tubes of a high-gain amplifier is obtained from a single source, the internal resistance of the source is common to all the plate circuits and is likely to act as a coupling between stages. Similar couplings may exist through a bleeder circuit when screen voltage for two or more tubes is taken from a common tap or through a bias resistor common to the control-grid circuits of several tubes. To reduce such stray couplings to negligible amounts, decoupling filters are generally inserted in the circuits of each tube and separate bias resistors are used.

FIG. 50. RC filter usage to isolate power-supply system from alternating currents.

A typical application of decoupling filters is shown in Fig. 50, the filter elements being indicated by heavy lines. Capacitors C furnish low-impedance paths back to the cathodes for the signal currents flowing in the grid, screen-grid, and plate circuits, while the high-impedance resistors R and chokes in the leads to the voltage divider prevent any appreciable flow of signal currents in that direction. The choice of values for these resistors and chokes depends principally upon the currents in the leads and the permissible d-c voltage drop in each filter. The impedance of each by-pass capacitor should be not more than 10 per cent of that of the associated resistor or choke, at any frequency for which the amplifier is designed to operate. On the other

FIG. 51. CR circuit for obtaining out-of-phase voltages.

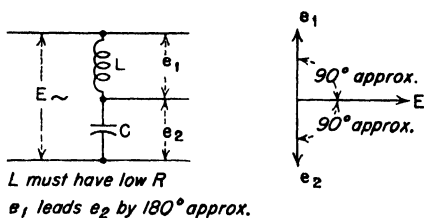



FIG. 52. LC circuit for obtaining out-of-phase voltages.

hand, the value of C should not be so large in any filter that "blocking" or motor-boating occurs due to too high a time constant.

The value of each cathode resistor, when separate biasing resistors are used, is equal to the bias required, divided by the total cathode d.c. of that tube. The screen-grid filter resistors serve as voltage-dropping resistors as well as filters, and their values are determined by the IR drops required for correct screen voltages.

31. Circuits for Obtaining Out-of-phase Voltages and Currents. Two circuits producing voltages 90 or 180 deg out of phase are shown in Figs. 51 and 52 with their vector diagrams. These are often useful in circuit designs and oscillograph measure-

ments. To maintain these phase relations, high-impedance circuits only should be connected across e_1 and e_2.

A circuit for obtaining currents 90 deg out of phase with each other is shown in Fig. 53.

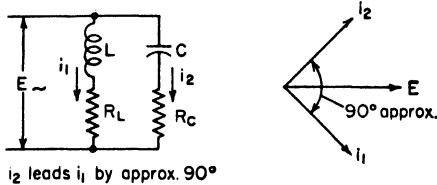



FIG. 53. Circuit for obtaining currents out of phase by 90 deg.

To utilize these currents, nonreactive loads R_L and R_C are introduced, with values such that $R_L R_C = L/C$ and $R_L = R_C$.

32. Frequency Discriminator Circuit. The frequency discriminator circuit shown in Fig. 54 is applied in a.f.c., f-m detection, frequency-drift indicators, etc. L_1C_1 and L_2C_2 are tuned to the same frequency and doubly coupled: (1) directly at B and (2) inductively by M. After rectification, a bias E is obtained which, between limits C and D, is proportional to the difference between the frequency of the input voltage and the resonance frequency of LC. The time constant of R_0C_0 should be much less than the period of one cycle of the *frequency variation* in the input voltage.

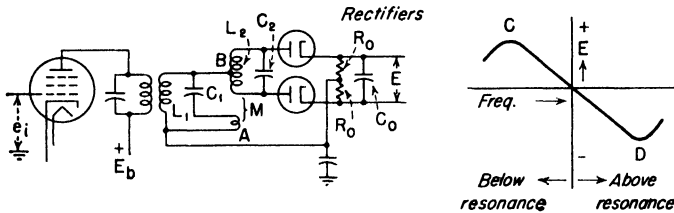



FIG. 54. Frequency discriminator and characteristic.

33. Compensation in Resistance-coupled Amplifier. In a conventional resistance-coupled amplifier (Fig. 55) the amplification falls off at low frequencies because of increasing impedance of C_c and at high frequencies because of the shunting effect of stray capacitance C_s. In wide-band amplifiers, the compensating impedances L and R_1C_1 are added. For approximately constant gain between frequency limits f_1 (low) and f_2 (high),

$$\left.\begin{aligned} R_L &= \frac{1}{2\pi f_2 C_s} \qquad & L &= \frac{R_L}{4\pi f_2} \\ C_1 &= \frac{1}{2\pi f_1 R_L} \qquad & R_1 &= 2R_L \end{aligned}\right\} \tag{97}$$

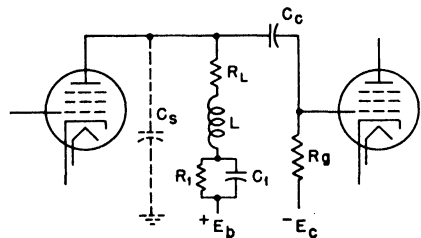



FIG. 55. Compensated resistance-coupled amplifier.

This type of compensation also tends to correct for phase shift near the limits f_1 and f_2.

NETWORKS

34. Types of Networks. A network is any combination of L, C, R, or M impedance elements having two or more terminals for connection to an external circuit. The eventual function of a network is to serve as part or all of a path joining a load to a source, but often the network is considered alone and the remainder of the circuit, including source and load, is merely implicit. If the network contains reactance, its effect on load current will usually vary with frequency; this is expressed quantitatively as the *response* or *characteristic* of the network, usually in the form of a curve or equation.

When the purpose is primarily to convey energy from source to load, the network is a *transmission line.* In this case the impedances are ordinarily distributed *LCRs* whose existence and values are more or less incidental to other factors, but these impedances may, nevertheless, play important roles in the response of the transmission line.

When the principal object is to control or alter the response of a system, the network is a *corrective* device, with deliberately planned impedances. At low frequencies, the impedances are usually lumped because relatively large *LCR* values are required, but at high frequencies they may consist of resonant sections of transmission line.

Depending upon the geometric arrangement of their elements, the most used networks are ladder, bridged-T, or lattice types. The basic sections of these networks

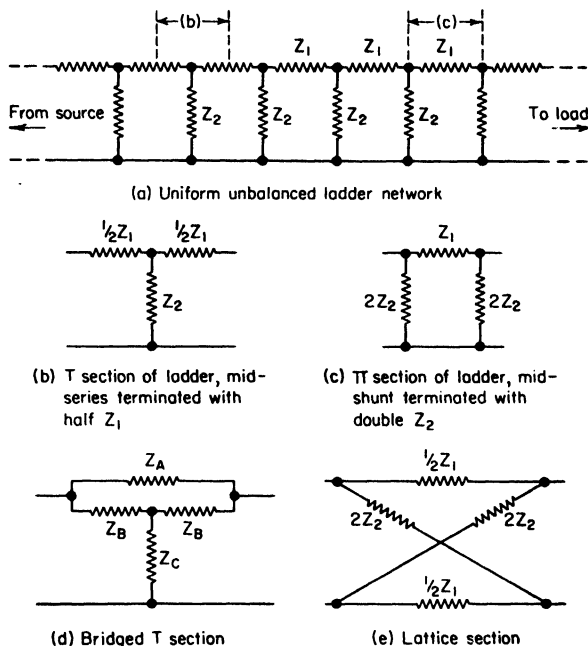



FIG. 56. Basic types of networks.

are shown in Fig. 56. Full series impedance of a ladder network is represented by Z_1, and full shunt arms by Z_2. In a single section removed from a ladder structure, these are halved or doubled as shown in Fig. 56*b* and *c*, so that connection with adjacent sections maintains the values Z_1 and Z_2.

It should be noted that current reaches the load through only one path in a ladder network, while in the bridged-T and lattice types there are two paths from input to output.

35. Poles and Zeros of Impedance. The zero impedance of a network or impedance branch at resonance is called a *zero,* and the infinite impedance at antiresonance frequency is called a *pole.* These expressions are frequently used in describing the characteristics of wave filters and equalizers, where the distribution of zeros and poles in the frequency spectrum is the factor that determines the frequency response of the structure.

Dissipative structures can have poles and zeros only at zero or infinite frequency. Absolute poles and zeros at all other frequencies occur only with pure reactance

networks. However, the terms are often applied when only incidental dissipation exists, as is the usual case in physical networks.

For locating poles and zeros in the frequency spectrum, ω ($= 2\pi f$), sometimes called the radian frequency, is generally used in place of frequency itself, to eliminate the constant repetition of the 2π factor.

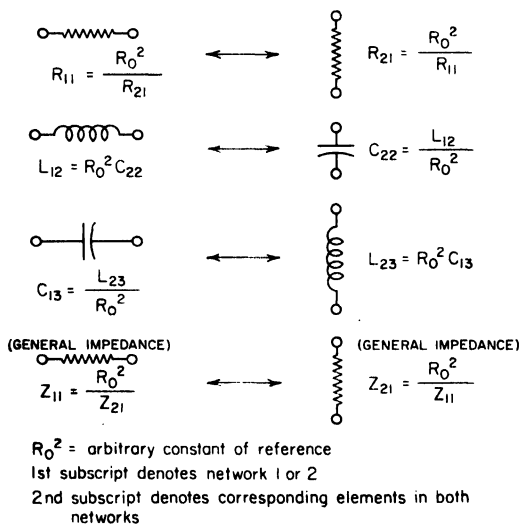
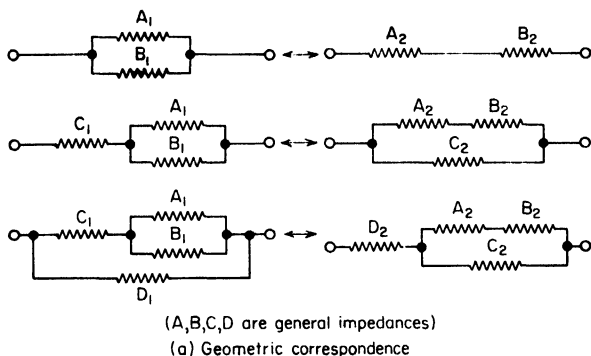



FIG. 57. Relations between two-terminal networks mutually inverse with respect to R_0^2.

36. Inverse Two-terminal Networks. Two networks whose impedance product is a constant are inverse with respect to the constant. That is, $Z_{11}Z_{21} = R_0^2$, where Z_{11} and Z_{21} are the network impedances and R_0^2 is the constant. Inverse two-terminal networks have direct application as the series and shunt branches of four-terminal constant-resistance structures. It should be noted that some networks (more complex than any considered here) do not have physically realizable inverse networks.

Definite relationships exist between each impedance element in a network and the corresponding element in the inverse network, with respect to geometrical arrange-

ment or method of connection to other elements, type of element (L, C, or R), and quantitative values. Geometrically, the inverse of a series element is one connected in parallel, and vice versa, as illustrated by Fig. 57*a*. An R in one network is represented by an R, usually of different value, in the inverse structure; an L in one corresponds to C in the inverse, and a C to an L, as shown by Fig. 57*b*. Quantitative relations between elements are also shown in the latter figure, and these may be summarized as

$$Z_{11}Z_{21} = R_{11}R_{21} = \frac{L_{11}}{C_{21}} = \frac{L_{12}}{C_{22}} = \frac{L_{21}}{C_{11}} = \frac{L_{22}}{C_{12}} = R_0^2 \tag{98}$$

In the conventional system of subscript notation used with inverse networks and impedances, the first integer represents the branch of the network to which an element

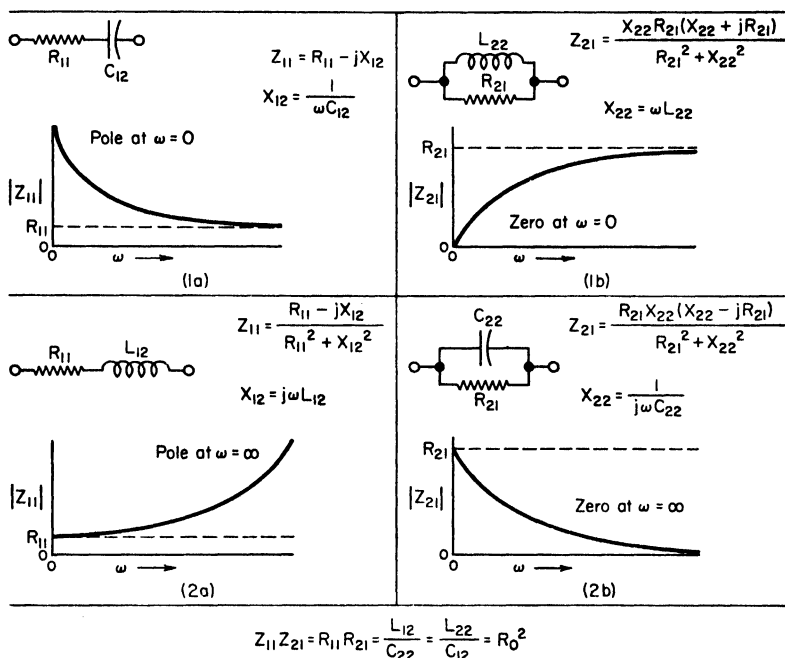



FIG. 58. Two-terminal networks mutually inverse with respect to R_0^2.

belongs, and the second integer designates a particular element in the branch. Thus, L_{12} denotes element number 2 in branch 1, and C_{22} the corresponding inverse element in branch 2.

At each frequency where branch 1 has a pole, its inverse branch 2 will have a zero, and vice versa.

A useful practical property of inverse networks is the fact that any two-terminal network inserted in series between two terminating resistances R_1 and R_2 such that $R_1R_2 = R_0^2$ produces exactly the same insertion loss as does its inverse with respect to R_0^2, when the latter is shunted across the circuit. Use is made of this principle in the design of simple equalizers.

Mutually inverse networks of two-, three-, and four-element types, with their respective equations and curves, are shown by Fig. 58. Actually, these are not all unique networks, since, for example, networks 3a and 4b are potentially equivalent as evidenced by the similarity of their impedance curves. Similar conditions apply

FIG. 58. Two-terminal networks mutually inverse with respect to R_0^2. (*Continued.*)

to 3b and 4a, 5a and 6a, and to 5b and 6b. Thus, only two impedance characteristics are obtainable with any possible combination of three reactance elements in a two-terminal network, and the two corresponding networks are inverse. The same situation exists with respect to four-element networks. This limits the types of response characteristics that may be achieved by employing such two-terminal networks as series and shunt branches in equalizers and wave filters.

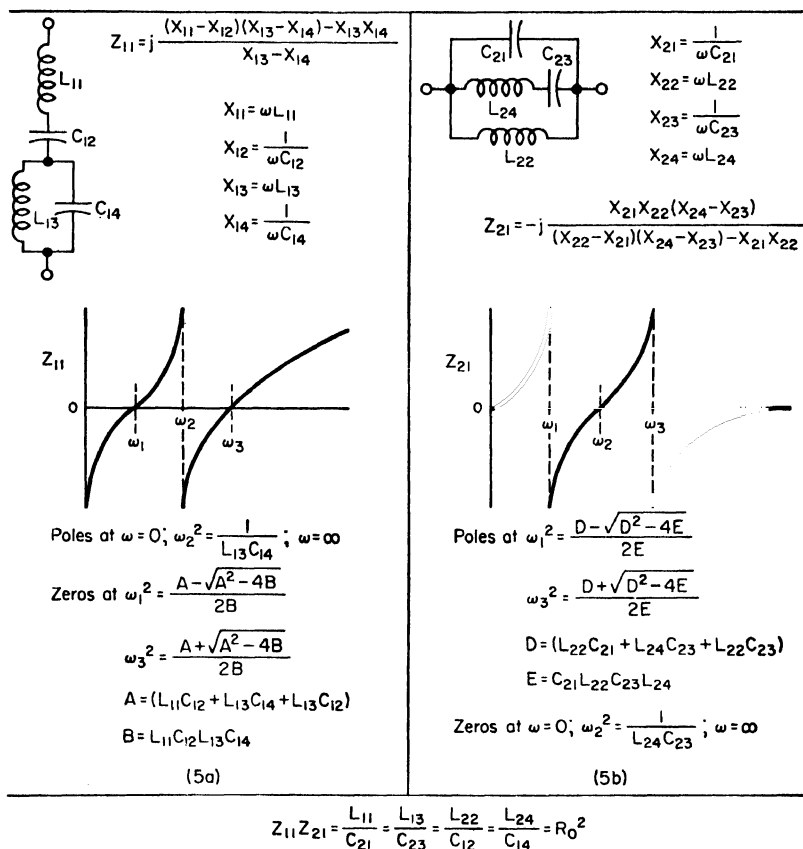



FIG. 58. Two-terminal networks mutually inverse with respect to R_0^2. (*Continued.*)

37. Lattice Networks. A symmetrical balanced lattice is shown in Fig. 59. To simplify drawing, the abbreviated form *b* is customarily used, where the dashed lines indicate series and cross impedances identical, respectively, to those explicitly shown. In the balanced lattice section, the total series impedance Z_{11} is equally divided between the two series branches as $\frac{1}{2}Z_{11}$, and each cross impedance Z_{21} is doubled to present mid-shunt termination. Impedances Z_{11} and Z_{21} represent any combinations of *L*, *C*, and *R*, although in most practical cases they are mutually inverse impedances.

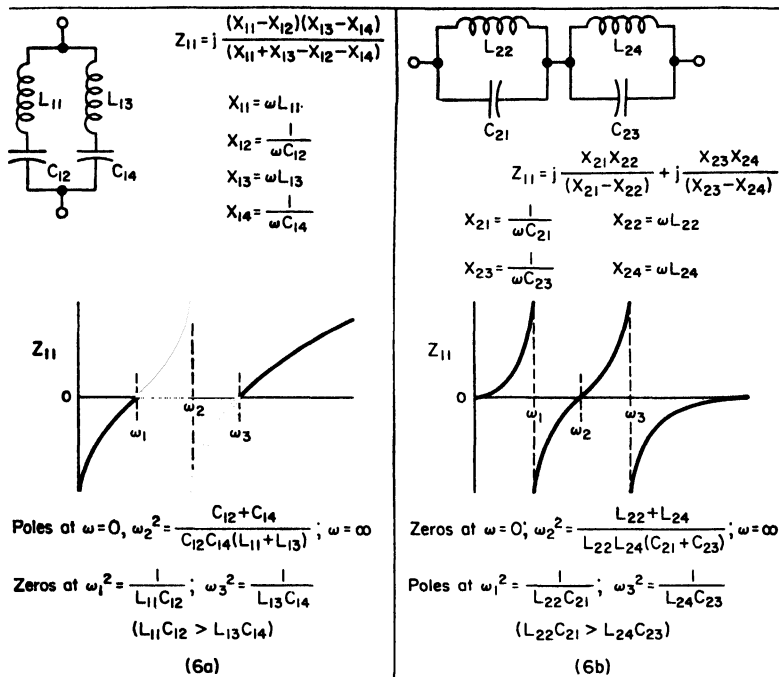
A lattice structure is more general than a bridged-T or ladder network, since it is physically realizable for some propagation characteristics for which no realizable bridged T or ladder exists. A physically realizable structure is one that can be built of real *L*s, *C*s, and *R*s (and mutual inductances in some cases), excluding all negative impedance values.

$$Z_{11}Z_{21} = \frac{L_{11}}{C_{21}} = \frac{L_{13}}{C_{23}} = \frac{L_{22}}{C_{12}} = \frac{L_{24}}{C_{14}} = R_0^2$$

FIG. 58. Two-terminal networks mutually inverse with respect to R_0^2. (*Continued.*)

The subscript notation used here for the lattice is the same as the system described for inverse networks in Sec. 36.

Practically, a lattice network has several disadvantages: (1) it usually requires more components than a bridged T or ladder; (2) it is more critical with respect to exact values of the elements and their adjustment; (3) it has no completely unbalanced form allowing one side to be grounded at both terminals.

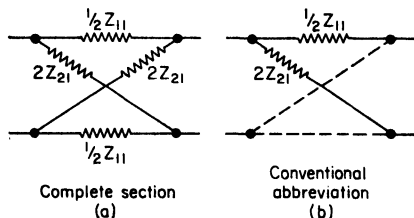



FIG. 59. Basic lattice section.

Offsetting these disadvantages is the fact that a lattice may be physically realizable for a desired attenuation or phase characteristic when other structures are not. Where it is necessary to use a lattice in an unbalanced circuit, transformers may be used for isolation if suitable ones are available.

38. Transformation of Lattice to Bridged T. Bridged-T equivalent networks may be derived from some lattices, although it is not possible in the case of every lattice. The bridged T has the advantage of requiring fewer components, and also may be used in its unbalanced form in circuits where a lattice would require isolation transformers. A lattice and one form of potentially equivalent bridged T are shown in Fig. 60*a* and *b*.

A method of lattice-to-bridged-T conversion described by Zobel[1] yields a physical bridged T when the lattice series branch Z_{11} includes a totally shunting resistance such as $\frac{1}{2}R_{13}$ in Fig. 60c. In this method, a hypothetical negative resistance $-2R_0/c$ is paralleled across the Z_{11} branch and made numerically equal to R_{11} as in Fig. 60d. Because the bridged T is unbalanced, the full value of R_{13} is used, instead of its half

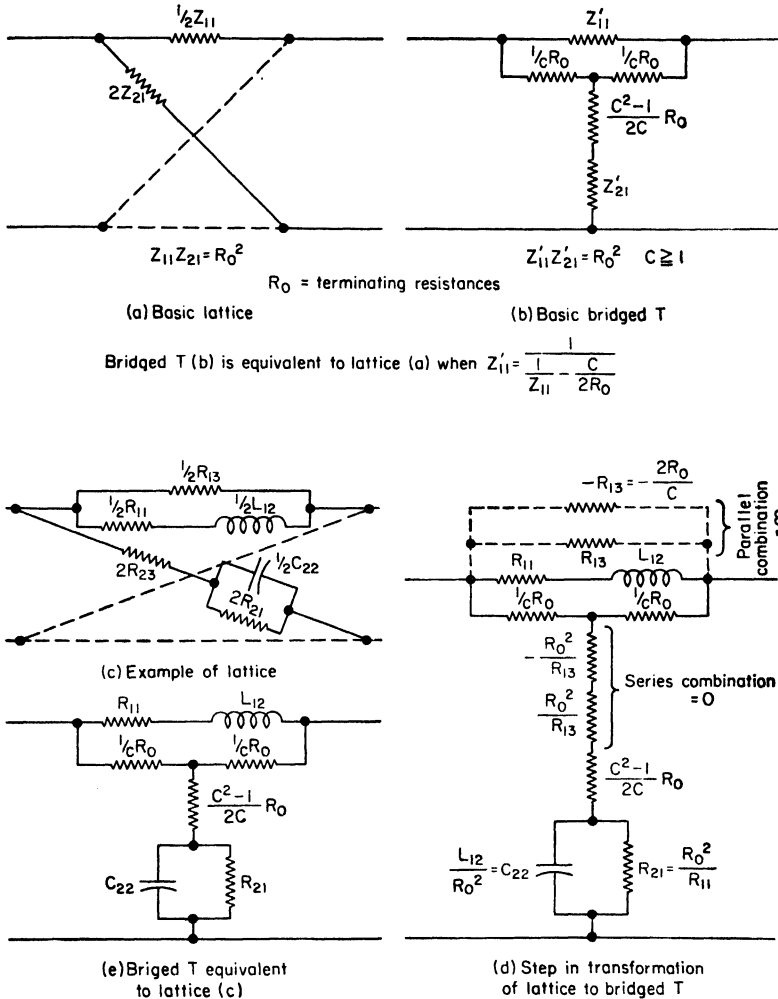



FIG. 60. Bridged T equivalent to lattice equalizer.

value as in the case of the lattice. The actual resistance R_{13} combined with the theoretical negative resistance $-2R_0/c$ is equivalent to infinite resistance; so both may be omitted. The value of c required to achieve this is $c = 2R_0/R_{13}$.

In the shunt branch of the bridged T, the inverse impedances of $-2R_0/c$ and R_{13}—namely, $-R_0^2/R_{13}$ and R_0^2/R_{13}—are in series, and these also effectively cancel

[1] Zobel, Otto J., Distortion Correction in Electrical Circuits with Constant Resistance Recurrent Networks, *Bell System Tech. J.*, **7** (3), 438, July, 1928.

each other. Thus they too can be omitted. Additional real resistances, $(1/c)R_0$ in the series branch and $[(c^2 - 1)/2c]R_0$ in the shunt branch, complete the bridged T as shown in Fig. 60*e*.

Of the lattice equalizers considered in Sec. 41, the types 1b, 2a, and 3a have the requisite shunt resistances in their series branches and therefore are convertible to bridged Ts by this method. The appropriate values for c in each of these types are given in the equalizer-design data.

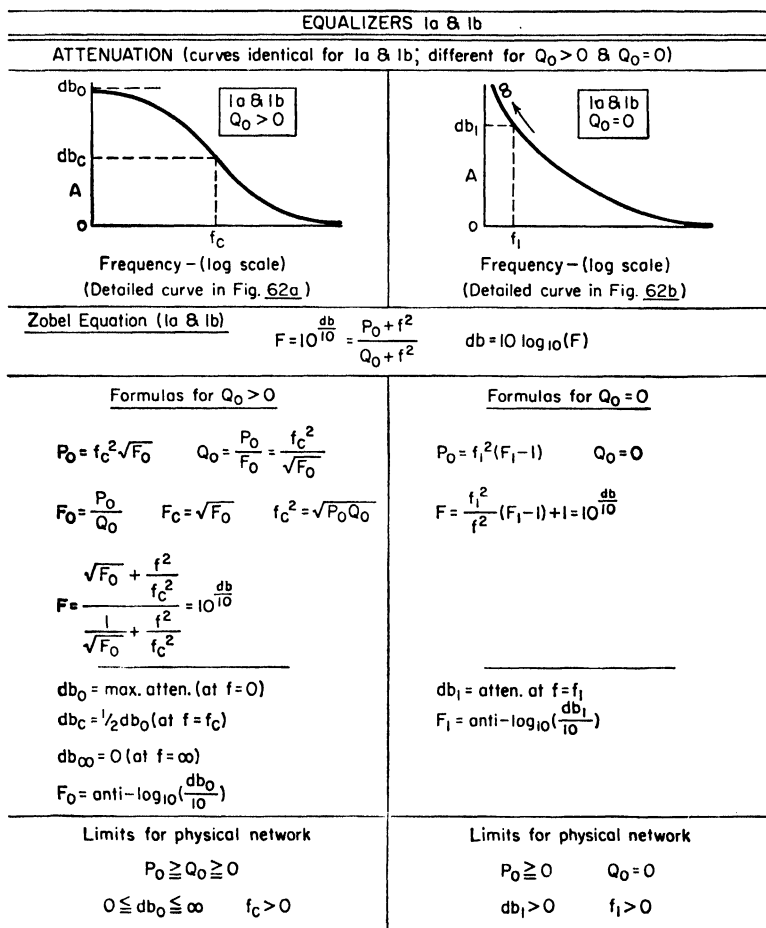



FIG. 61*a*. Equalizers.

Other methods exist for the conversion of lattices to bridged Ts but are too lengthy to be included here. These may be found in the works of Zobel, Bode, Guillemin, and others.

39. Constant-resistance Networks. A section of recurrent network having full series impedance of Z_{11} and full shunt impedance of Z_{21} is of constant-resistance type when $Z_{11}Z_{21} = R_0^2$. Its iterative impedance is equal to R_0. It follows that, when

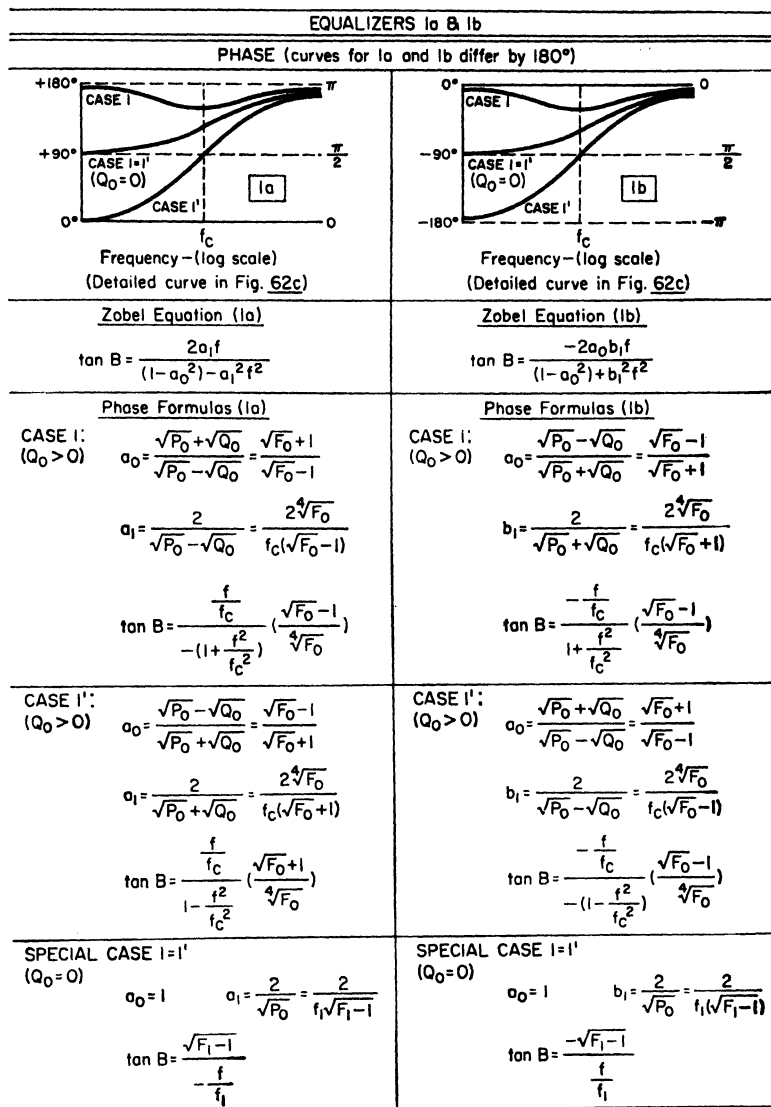



FIG. 61b. Equalizers.

the output terminals of a constant-resistance network are terminated in a resistive load R_0, the input impedance also equals R_0 at all frequencies.

The impedance Z_{11} for the types of networks considered in this chapter is limited to combinations of L, C, and R. In a constant-resistance structure, Z_{11} and Z_{21} are inverse two-terminal networks, with respect to R_0^2 (see Sec. 36).

Constant-resistance networks have advantages in communications circuits, particularly when used as equalizers, phase correctors, and predistorting devices. There

is usually some point in a communications system where the impedance looking toward the load is resistive and equal to the impedance looking toward the source, *i.e.*, "matched." At any such point, one or more constant-resistance networks of identical resistance may be inserted without disturbing the matched condition. This avoids reflection and interaction losses, so that insertion loss and phase shift introduced by the network are simply obtained as the sum of its propagation constant $P = A + jB$ and that of the original circuit. Thus, when a properly terminated constant-resistance network is employed as a corrective network, its effect upon transmission is completely described by its attenuation constant A and its phase constant B. These constants form the basis for the design of the device.

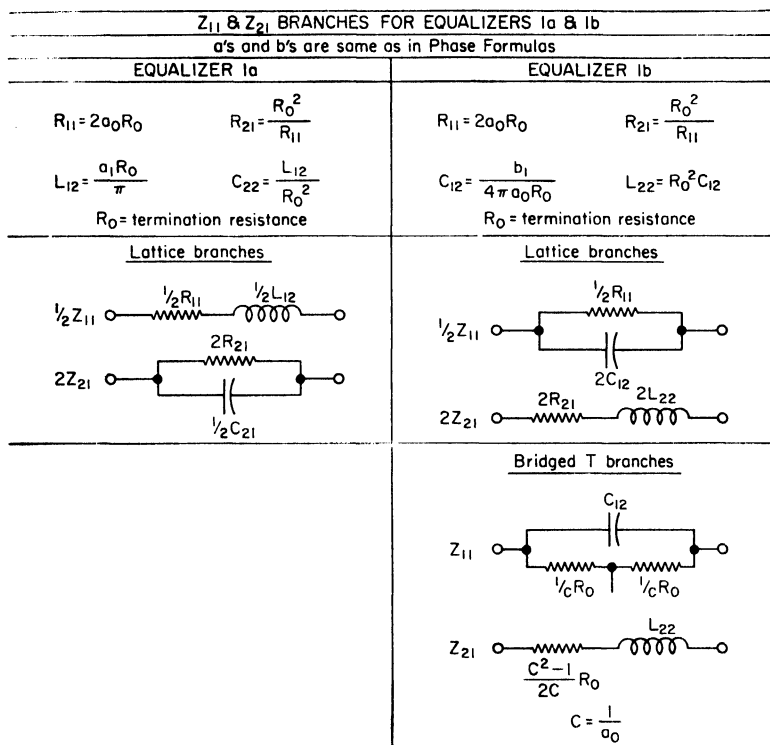



FIG. 61*c*. Equalizers.

40. Equalizers. An equalizer is a corrective network whose principal function is to alter the attenuation vs. frequency response of a communications circuit. Most equalizers have also an inherent phase characteristic that is by nature associated with their attenuation-frequency curves. Hence, if both equalization and phase correction are to be applied to a circuit, the equalizer should be designed first so that its contribution to the total phase shift can be included in the phase corrector.

As a basic unit for design purposes, the constant-resistance lattice is convenient. Some lattices—but not all—are convertible into constant-resistance bridged Ts for use in unbalanced circuits. The constant-resistance feature readily permits two or more sections having different characteristics to be joined in tandem, to achieve composite effects not obtainable with a single section.

One method of finding a network that will produce the required attenuation and phase characteristics consists of selection by comparison with a number of structures whose characteristics have been previously computed and plotted. From a reasonably complete list of this sort, equalizers for many ordinary applications may be designed. This is the basis of the equalizer-design data in Sec. 41. When more exacting specifications have to be met, the techniques of network synthesis are employed. These methods are too lengthy a subject to be included in a general handbook but are available in the works of Bode, Guillemin, and others.

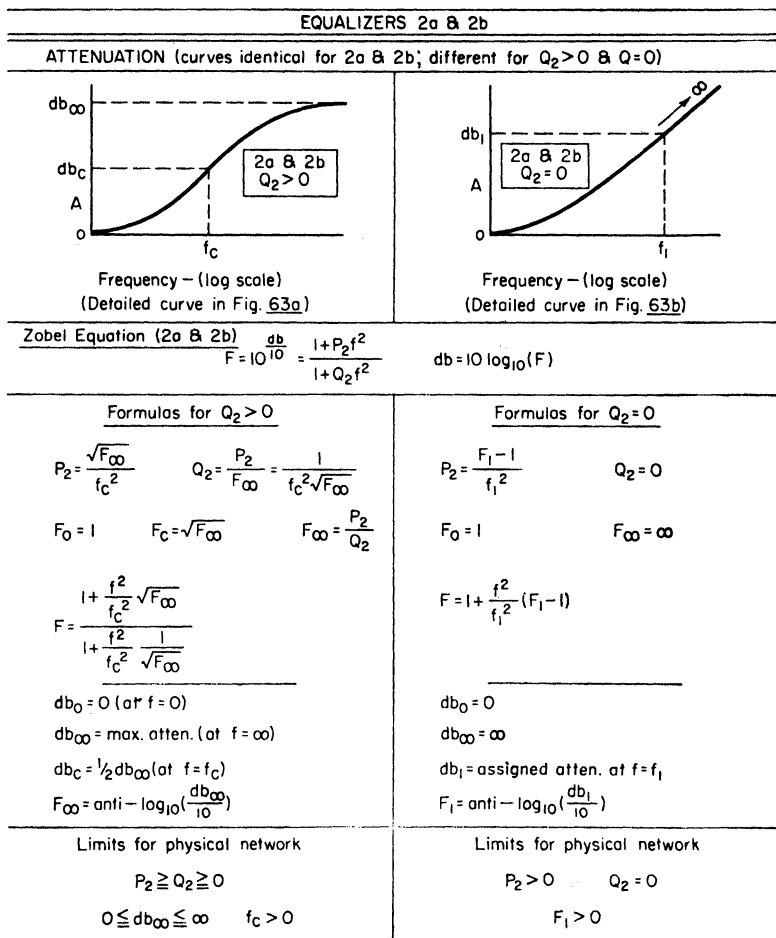



FIG. 61*d*. Equalizers.

In an excellent analysis of constant-resistance equalizers and phase correctors, Zobel[1] derived linear equations of the form

$$F \equiv e^{2A} = 10^{\frac{db}{10}} = \frac{P_0 + P_2 f^2 + \cdots}{Q_0 + Q_2 f^2 + \cdots}$$

[1] *Ibid*

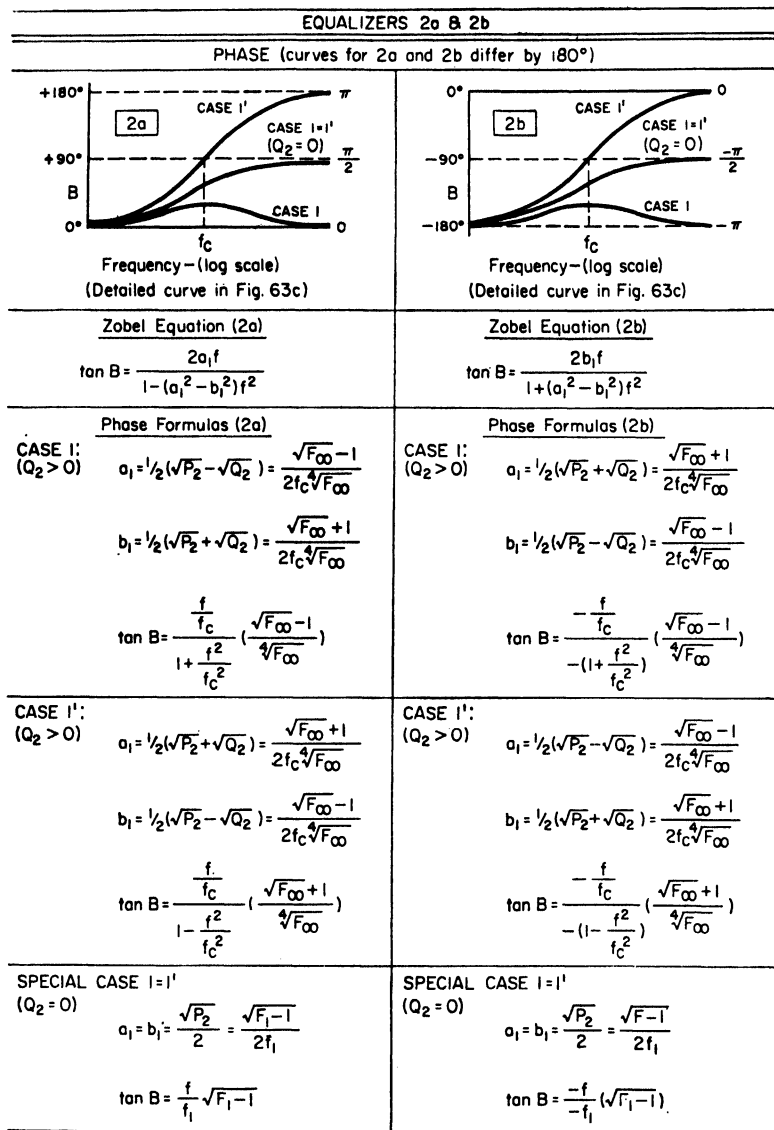



FIG. 61*e*. Equalizers.

for a number of explicit lattice networks. *F* is an amplitude-transfer factor, *A* the attenuation constant, db the attenuation; the *P*s and *Q*s are constants for a given lattice under given conditions, and *f* is the frequency in cycles. Substitution of pairs of values for attenuation and frequency according to the characteristic desired yields a set of equations to be solved first for the *P*s and *Q*s, and eventually for the lattice impedances in terms of *L*, *C*, and *R*.

In some cases, more than one set of L, C, and R values will satisfy the equations, thus permitting some choice of the attenuation and phase characteristic. In the design data of Sec. 41, these conditions are designated as Case 1, Case 1′, etc.

Certain of Dr. Zobel's results are used here, by permission, in the sections on equalizer and phase-corrector design, and in the lattice-to-bridged-T transformation of Sec. 38.

41. Equalizer-design Data. The tables in Fig. 61 list the Zobel equations for attenuation and phase shift of four constant-resistance lattice equalizers. In addi-

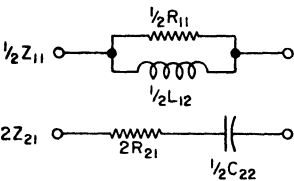
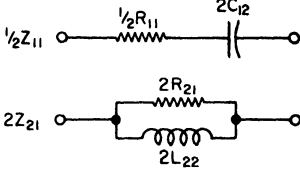
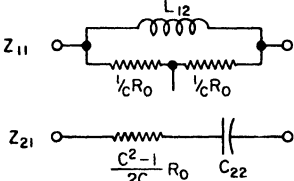



FIG. 61*f*. Equalizers.

tion, alternative relations are given, in terms of maximum and half-maximum attenuation. The latter are derived from the Zobel equations and offer somewhat shorter design formulas, but at the expense of some freedom of choice of points on the attenuation-frequency curves.

The tables also show the Z_{11} and Z_{21} branches for the lattice section (Fig. 60*a*) and for bridged-T equivalents (Fig. 60*e*) where they exist. Detailed attenuation and phase curves are shown in Figs. 62 to 65.

Space limitations prevent the inclusion here of more than these simpler forms of equalizers that are nevertheless useful in many ordinary applications. For more critical requirements and more complete techniques in the art of equalizer design, reference should be made to the works of Bode, Guillemin, etc.

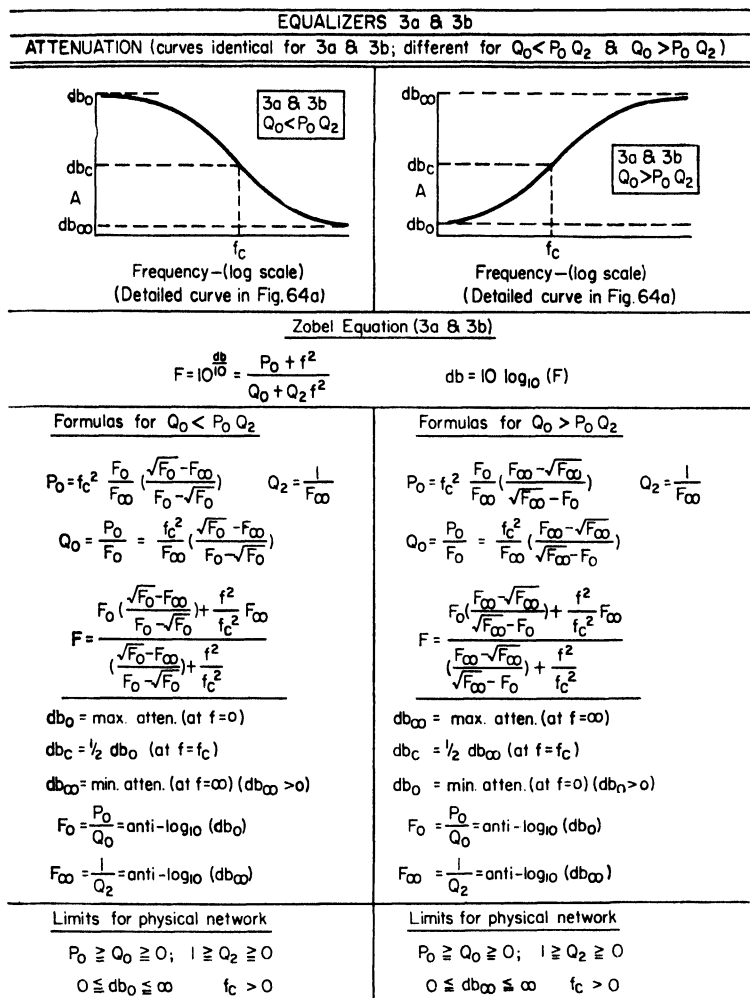



FIG. 61*g*. Equalizers.

To use the design data in this section, after the attenuation characteristic required of the equalizer has been determined:

1. Select from the table an attenuation curve that most nearly fits the required characteristic. (Note that the symbolic curves of the table are plotted against log scales for frequency and therefore appear to be symmetrical; on a linear scale they are not symmetrical.)

2. Choose the *a* or *b* form of the equalizer, on the basis of phase curves. Attenuation curves from both forms are identical, but the phase characteristics differ by 180 deg. If a bridged-T final form is desired, one of the forms 1b, 2a, or 3a should be used; these can be transformed to physical bridged Ts.

3. Use the design formulas to compute branch impedance values.

4. Compute phase characteristic from the tan *B* formula. NOTE: To determine the quadrant of the phase angle, carry through the negative signs attached to numerator and

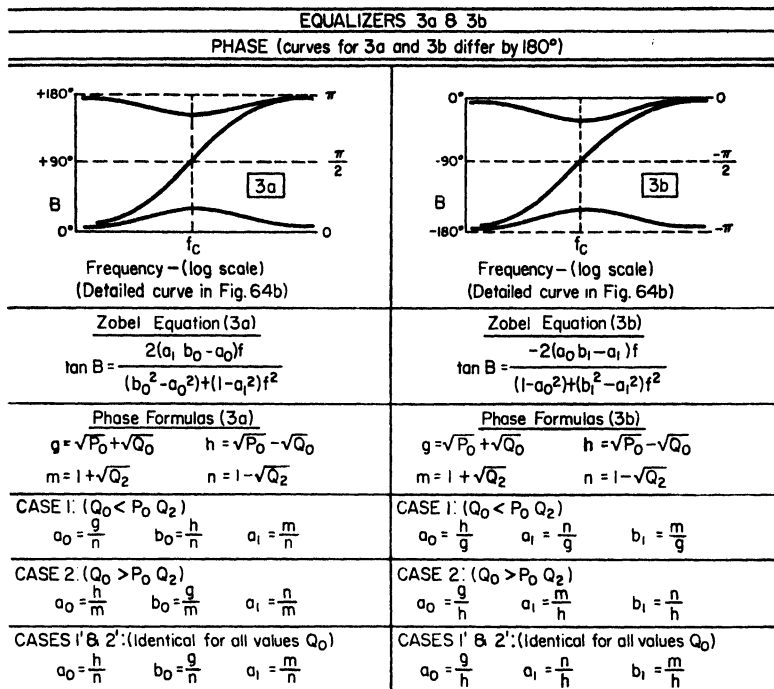



FIG. 61h. Equalizers.

denominator of the tan B equation to the final numerical result. The quadrant in which the angle lies can then be determined by the following table.

Numerator	Denominator	Quadrant
+	+	First
+	−	Second
−	−	Third
−	+	Fourth

Examples of Design of Equalizer:

1. Equalizer type 2a to have 4 db attenuation at 500 cycles, 10 db attenuation at 5,000 cycles, and $R_0 = 1{,}000$ ohms. Using Zobel attenuation equations,

$$F_{500} = 10^{0.4} = 2.512 = \frac{1 + P_2(500)^2}{1 + Q_2(500)^2}$$

$$F_{5,000} = 10^{1.0} = 10 = \frac{1 + P_2(5{,}000)^2}{1 + Q_2(5{,}000)^2}$$

These yield $P_2 = 7.956 \times 10^{-6}$ and $Q_2 = 0.7596 \times 10^{-6}$.

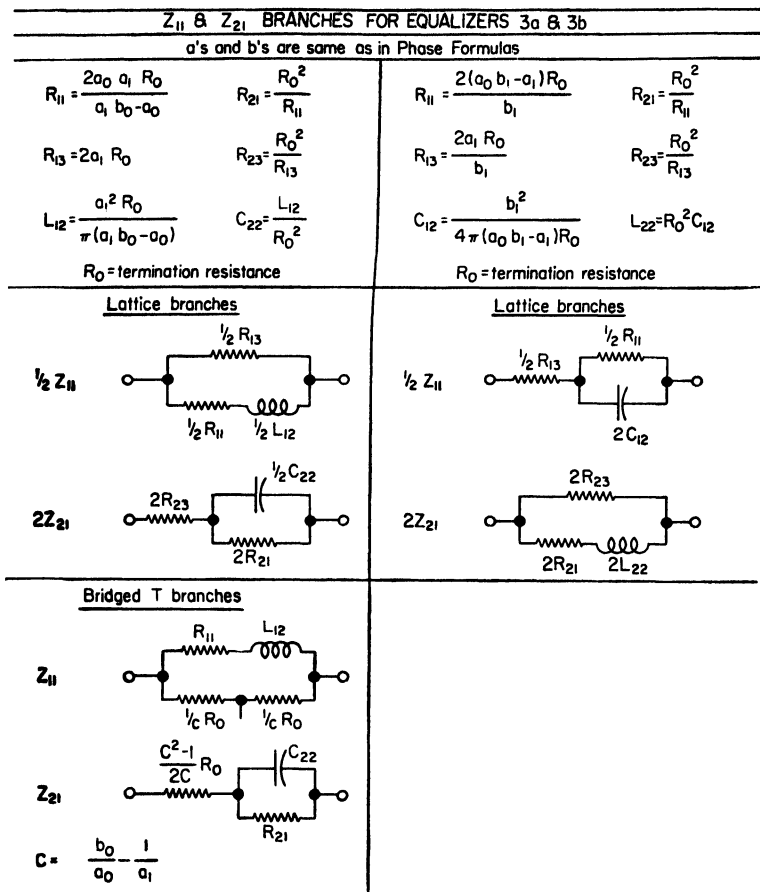



Fig. 61*i*. Equalizers.

Using Case 1 conditions:

$$a_1 = \tfrac{1}{2}(\sqrt{P_2} - \sqrt{Q_2}) = 0.9747 \times 10^{-3}$$

and

$$b_1 = \tfrac{1}{2}(\sqrt{P_2} + \sqrt{Q_2}) = 1.8463 \times 10^{-3}$$

$$R_{11} = \frac{2a_1R_0}{b_1} = 1{,}055.8 \text{ ohms}$$

$$R_{21} = \frac{R_0^2}{R_{11}} = 947.1 \text{ ohms}$$

$$L_{12} = \frac{a_1R_0}{\pi} = 0.3103 \text{ henry}$$

$$C_{22} = \frac{L_{12}}{R_0^2} = 0.3103 \times 10^{-6} \text{ farad}$$

$$\tan B = \frac{2a_1f}{1 - (a_1^2 - b_1^2)f^2} = \frac{1.9494f}{1 + 4.359f^2}$$

Numerator and denominator are both positive; so B lies in first quadrant for all positive values of f.

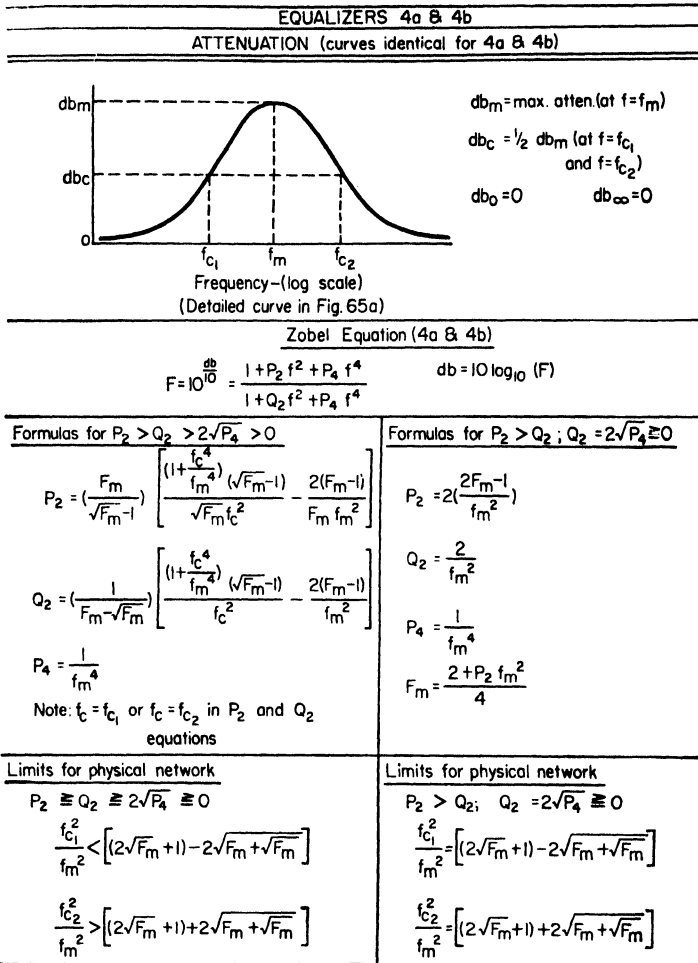



FIG. 61*j*. Equalizers.

Transformation to bridged T, in accordance with Sec. 38:

Let
$$\frac{2R_0}{c} = R_{11} \text{ of lattice} = \frac{2a_1R_0}{b_1}$$
$$c = \frac{b_1}{a_1} = 1.8942$$
$$\frac{1}{c}R_0 = 527.9$$
$$\frac{c^2 - 1}{2c}R_0 = 683.1$$

The completed lattice and bridged-T equalizers are shown in Fig. 66.

2. Equalizer type 2a to have 14 db maximum attenuation at $f = \infty$ and ½ maximum attenuation at $f_c = 1{,}000$ cycles, with $R_0 = 1{,}000$ ohms.

Using formulas for 2a, $Q_2 > 0$:

$$F_\infty = 10^{1.4} = 25.12$$

$$P_2 = \frac{\sqrt{F_\infty}}{f_c^2} = 5.012 \times 10^{-6}$$

$$Q_2 = \frac{P_2}{F_\infty} = 0.1996 \times 10^{-6}$$

The balance of the computation is similar to that of the first example.

42. Degenerate Forms of Equalizers. Several simplified versions of equalizers, requiring fewer elements to construct than the full-section constant-resistance types,

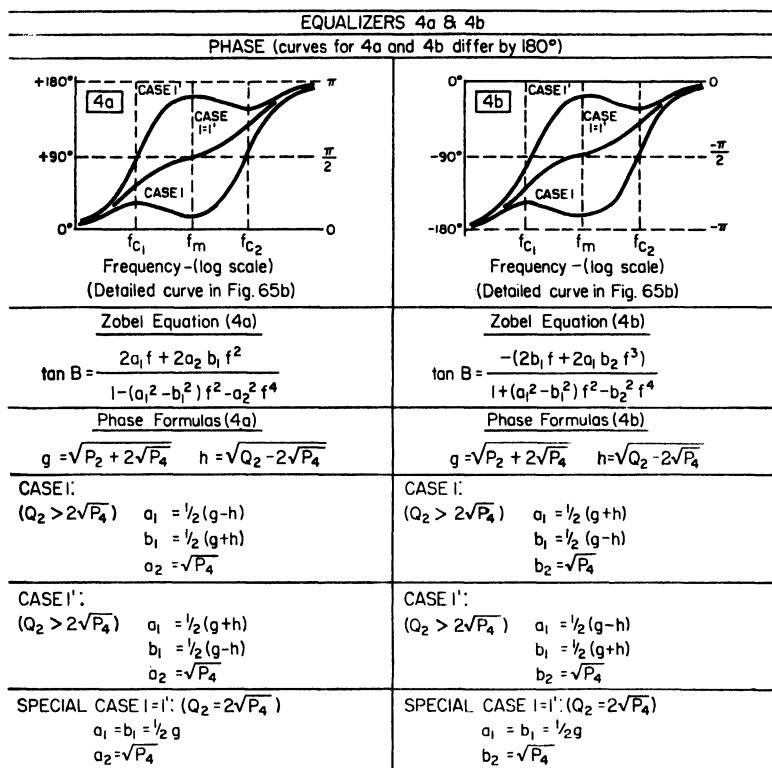



FIG. 61*k*. Equalizers.

are shown in Fig. 67. In each case, either the series or shunt reactance arms are omitted. The branches retained are identical with corresponding ones in the full lattice in *a* and *b*, or with those of a bridged T in *c* and *d*. Resistances equal to R_0 are substituted for the omitted branches in *a* and *b*, resulting in 6 db additional insertion loss for these as compared with a full lattice structure.

These degenerate forms have the same attenuation characteristics as their corresponding full-section lattices or bridged Ts, but only when used between terminations each exactly equal to R_0. Since they are not in themselves constant-resistance devices, they do not present proper resistive terminations for other networks, and

Z_{11} & Z_{21} BRANCHES FOR EQUALIZERS 4a & 4b			
a's and b's are same as in Phase Formulas			
EQUALIZER 4a		EQUALIZER 4b	
$R_{11}=\dfrac{2a_1^2R_0}{a_1b_1-a_2}$	$R_{21}=\dfrac{R_0^2}{R_{11}}$	$R_{11}=\dfrac{2(a_1b_1-b_2)R_0}{b_1^2}$	$R_{21}=\dfrac{R_0^2}{R_{11}}$
$L_{12}=\dfrac{a_1a_2R_0}{\pi(a_1b_1-a_2)}$	$C_{22}=\dfrac{L_{12}}{R_0^2}$	$C_{12}=\dfrac{b_1b_2}{4\pi(a_1b_1-b_2)R_0}$	$L_{22}=R_0^2C_{12}$
$L_{13}=\dfrac{a_1R_0}{\pi}$	$C_{23}=\dfrac{L_{13}}{R_0^2}$	$C_{13}=\dfrac{b_1}{4\pi R_0}$	$L_{23}=R_0^2C_{13}$
R_0 = termination resistance		R_0 = termination resistance	
Lattice branches		Lattice branches	
$\frac{1}{2}Z_{11}$: $\frac{1}{2}L_{13}$ in parallel with $\frac{1}{2}R_{11}$, $\frac{1}{2}L_{12}$		$\frac{1}{2}Z_{11}$: $2C_{13}$, $\frac{1}{2}R_{11}$, $2C_{12}$	
$2Z_{21}$: $\frac{1}{2}C_{23}$, $\frac{1}{2}C_{22}$, $2R_{21}$		$2Z_{21}$: $2L_{23}$, $2R_{21}$, $2L_{22}$	

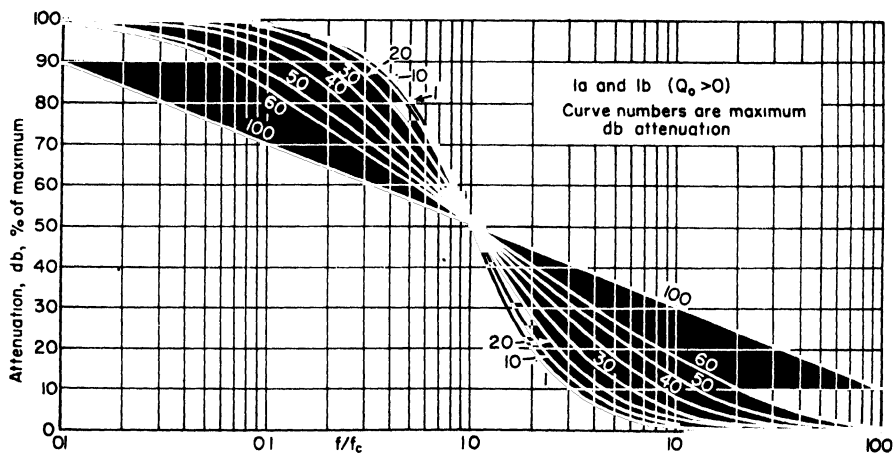



FIG. 61*l*. Equalizers.

FIG. 62*a*. Typical attenuation curves for equalizers 1*a* and 1*b*, $Q_0 > 0$.

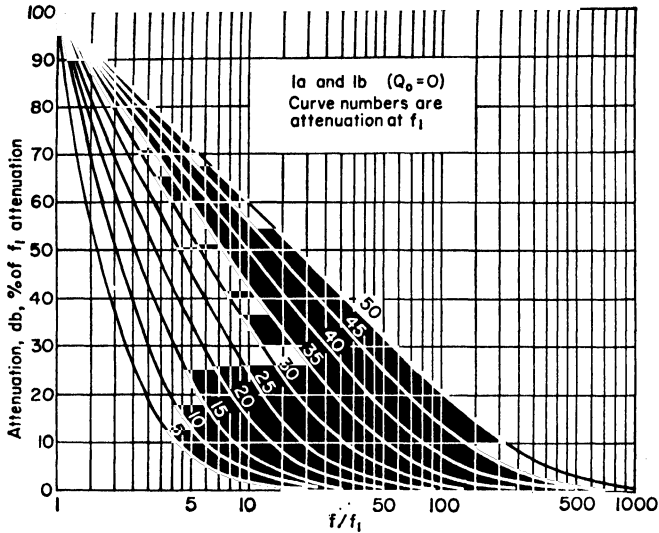



FIG. 62b. Attenuation curves for equalizers 1a and 1b, $Q_0 = 0$.

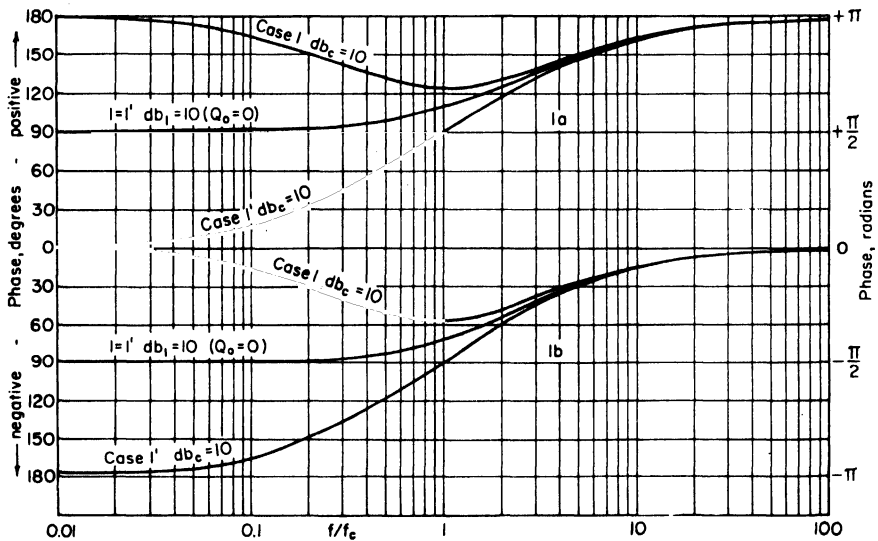



FIG. 62c. Typical phase curves for equalizers 1a and 1b.

this precludes their use in tandem with other sections. Their chief advantage therefore lies in their economy for use as single sections in simple applications.

43. Phase Correctors. A constant-resistance lattice whose branches are pure reactances theoretically has no attenuation at any frequency, so is an all-pass network. The absence of attenuation while at the same time a phase characteristic is retained makes this type of network useful for phase correction. Data on several lattices of this type were published by Zobél,[1] and his equations for the following two forms are reproduced here by permission.

[1] *Ibid.*

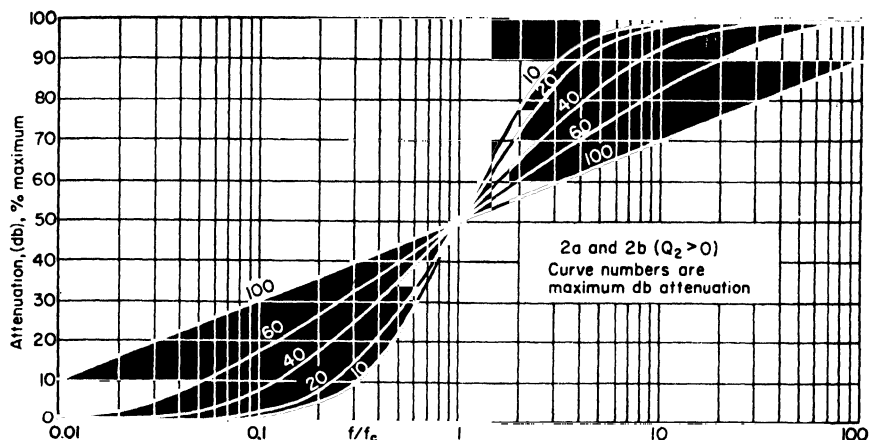



FIG. 63a. Typical attenuation curves for equalizers $2a$ and $2b$, $Q_2 > 0$.

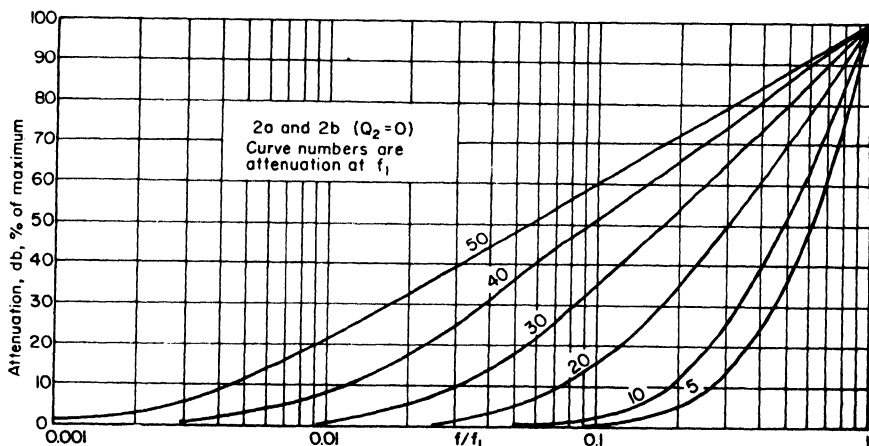



FIG. 63b. Typical attenuation curves for equalizers $2a$ and $2b$, $Q_2 = 0$.

Phase Corrector I. The lattice circuit is shown in Fig. 68, and a typical phase curve in Fig. 69. The basic equations are:

$$\tan \tfrac{1}{2} B = a_1 f \tag{99}$$

$$L_{11} = \frac{a_1 R_0}{\pi} \tag{100}$$

$$C_{21} = \frac{L_{11}}{R_0^2} = \frac{a_1}{\pi R_0} \tag{101}$$

where R_0 = source and load resistances (equal)
B = phase difference, output vs. input
a_1 = constant to be assigned, determining slope of phase characteristic
f = frequency, cps

In one sense, the shape of the phase curve for this lattice is fixed, since the slope is proportional to frequency for any given a_1. The design problem, then, is merely to

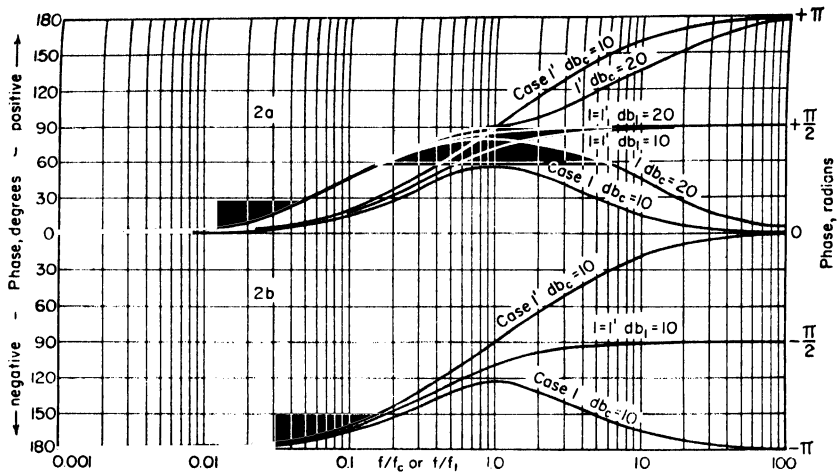



FIG. 63c. Typical phase curves for equalizers 2a and 2b.

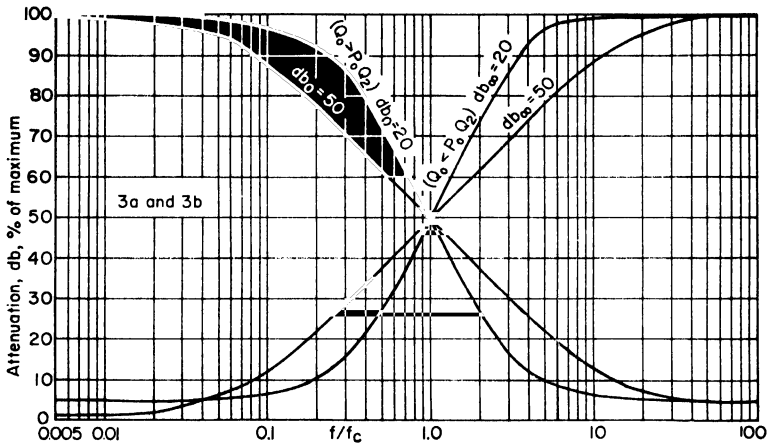



FIG. 64a. Typical attenuation curves for equalizers 3a and 3b.

evaluate a_1 to produce the desired phase angle at a specified frequency. This choice of a_1 then automatically determines the phase curve for all other frequencies.

Phase Corrector II. This lattice is shown in Fig. 70 and typical phase curves in Fig. 71. The basic equations are:

$$\tan \tfrac{1}{2} B = \frac{a_1 f}{1 - b_2 f^2} \tag{102}$$

$$L_{11} = \frac{a_1 R_0}{\pi} \tag{103}$$

$$C_{12} = \frac{b_2}{4\pi a_1 R_0} \tag{104}$$

$$C_{21} = \frac{L_{11}}{R_0^2} = \frac{a_1}{\pi R_0} \tag{105}$$

$$L_{22} = R_0^2 C_{12} = \frac{b_2 R_0}{4\pi a_1} \tag{106}$$

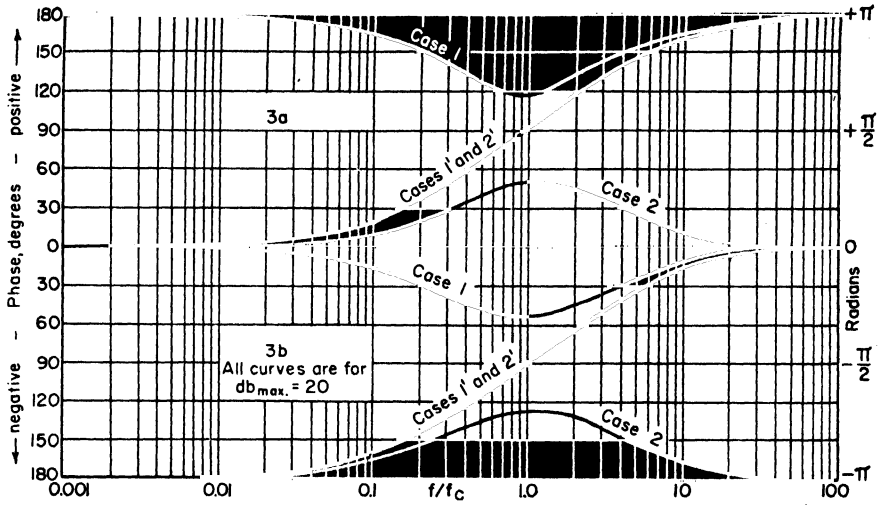



FIG. 64b. Phase curves for equalizers $3a$ and $3b$.

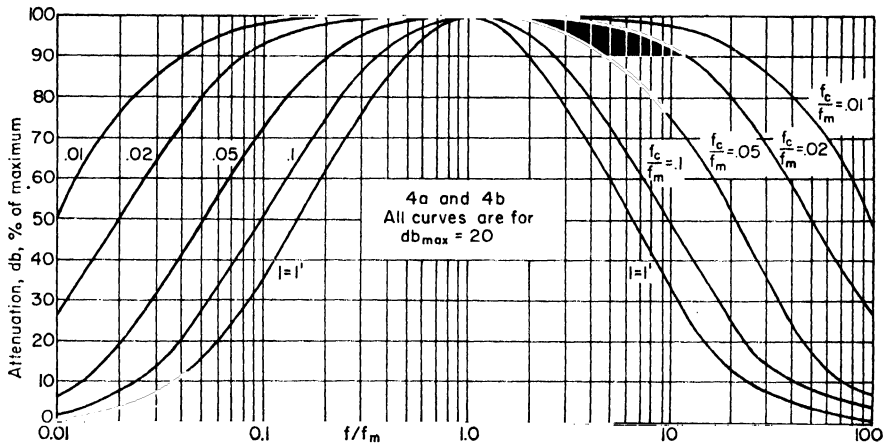



FIG. 65a. Typical attenuation curves for equalizers $4a$ and $4b$.

The symbols are the same as for type I, except for the additional constant b_2 to be assigned.

The presence of two arbitrary constants in Eq. (102) imparts some flexibility to the shape of the phase curve for this network. For design purposes, the substitution of two sets of values for phase angle and frequency into Eq. (102) will yield a pair of equations to evaluate a_1 and b_2. It should be noted that

$$f = \frac{1}{\sqrt{b_2}} \tag{107}$$

is the frequency at which the phase difference is 180 deg.

Conversion of all-pass lattices into bridged-T networks is discussed by Zobel, Bode, Guillemin, and others in the references, but the process involves additional techniques too lengthy to be included here.

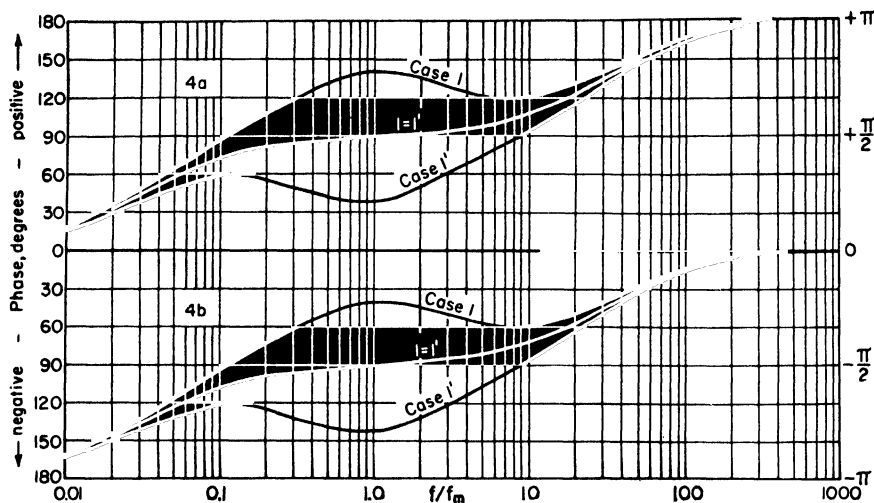



Fig. 65b. Phase curves for equalizers 4a and 4b.

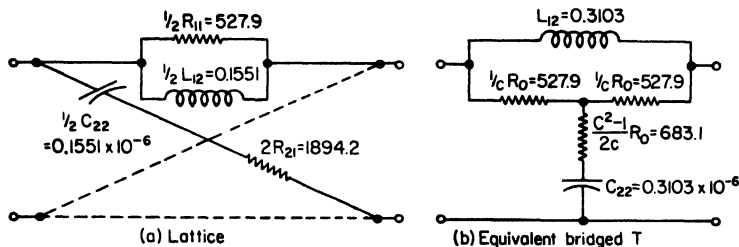



Fig. 66. Example of equalizer design.

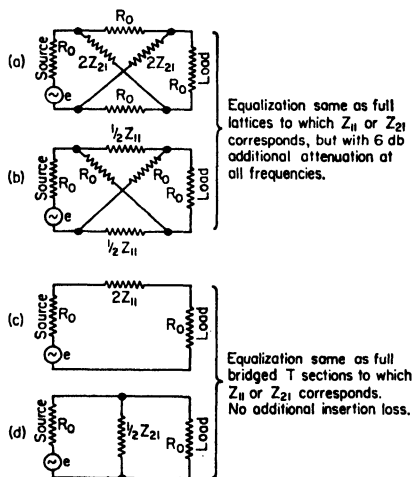
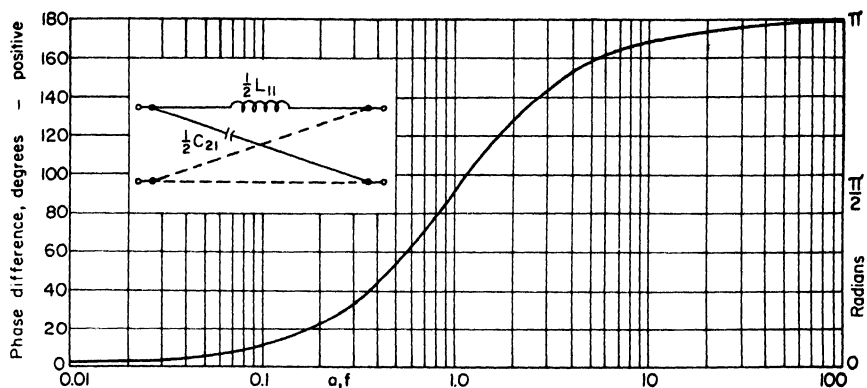
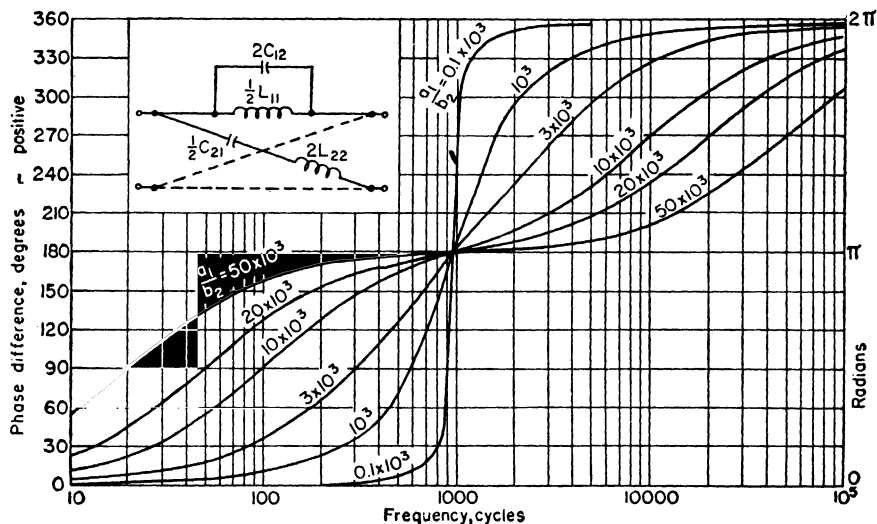



Fig. 67. Degenerate forms of equalizers.

FIGS. 68 and 69: FIG. 69. Typical phase response for type I phase corrector, Fig. 68 (insert).

FIGS. 70 and 71: FIG. 71. Typical phase response for type II phase corrector, Fig. 70 (insert).

TRANSMISSION LINES

44. General Properties of Transmission Lines. All the relations in this section apply to any smooth transmission line with uniformly distributed constants and are exact except that radiation and end effects are not taken into account.

Characteristic impedance is equivalent to the input impedance of an infinitely long line and is

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \quad \text{ohms} \tag{108}$$

or

$$|Z_0| = \sqrt[4]{\frac{R^2 + \omega^2 L^2}{G^2 + \omega^2 C^2}} \quad \text{ohms (magnitude)} \tag{109}$$

or

$$Z_0 = \sqrt{Z_{oc} Z_{sc}} \quad \text{ohms} \tag{110}$$

where Z_{oc} and Z_{sc} are the line input impedances with the far end open- and short-circuited, respectively, and R, L, G, and C are the resistance, inductance, leakage conductance, and capacitance of the line per unit of physical length. For practical purposes any convenient unit length such as 1 ft or 1 cm may be used. The same unit length must be used throughout any one computation, however.

The ratio of the currents I_1 and I_2 at points along the line one unit length apart is determined by the *propagation constant,* which is

$$P \text{ per unit length} = \log_\epsilon \frac{I_1}{I_2} = \sqrt{(R + j\omega L)(G + j\omega C)}$$
$$= A + jB \qquad (111)$$

Attenuation Constant. The real part A of P determines the relative magnitude of I_1/I_2 and is

$$A = 6.141 \sqrt{\sqrt{(R^2 + \omega^2 L^2)(G^2 + \omega^2 C^2)} + RG - \omega^2 LC} \quad \text{db/unit length} \qquad (112)$$

Wavelength Constant. The quadrature part B of P determines the relative phase of I_1/I_2 and is

$$B = 0.707 \sqrt{\sqrt{(R^2 + \omega^2 L^2)(G^2 + \omega^2 C^2)} - RG + \omega^2 LC} \quad \text{radians/unit length} \qquad (113)$$

To obtain the value in degrees per unit length, change the constant before the radical in Eq. (113) from 0.707 to 40.51.

Velocity of propagation is

$$V = \frac{\omega}{B} = \frac{2\pi f}{B} \quad \text{unit lengths/sec} \qquad (114)$$

Wavelength is

$$\lambda = \frac{2\pi}{B} \quad \text{unit lengths} \qquad (115)$$

Retardation time or *delay* is

$$t = \frac{B}{\omega} = \frac{B}{2\pi f} \quad \text{sec/unit length} \qquad (116)$$

Line Terminated in Any Impedance Z_R. The input impedance is

$$Z_i = \frac{E_s}{I_s} = Z_0 \frac{Z_R \cosh Pl + Z_0 \sinh Pl}{Z_0 \cosh Pl + Z_R \sinh Pl}$$
$$= Z_0 \frac{Z_R/Z_0 + \tanh Pl}{1 + (Z_R/Z_0) \tanh Pl} \quad \text{ohms} \qquad (117)$$

where l = length of line in unit lengths
P = propagation constant per unit length
Z_R = receiving end termination impedance
Z_0 = characteristic impedance of line
Z_i = input impedance of line
E_s and I_s = voltage and current at sending end

Current at any point x unit lengths from sending end in terms of sending end current is

$$I_x = I_s \left(\cosh Px - \frac{Z_i}{Z_0} \sinh Px\right) \quad \text{amp} \qquad (118)$$

Voltage at any point x unit lengths from sending end in terms of sending end voltage is

$$E_x = E_s \left(\cosh Px - \frac{Z_0}{Z_i} \sinh Px\right) \quad \text{volts} \qquad (119)$$

In this case, the current and voltage along the line vary from point to point because of standing waves. (See Sec. 58 for conditions in r-f lines.)

Lines Terminated in Impedance Z_0. The input impedance for any length of line is

$$Z_i = Z_0 \quad \text{ohms} \tag{120}$$

Current or voltage at any point x unit lengths from sending end in terms of sending end current or voltage is

$$\frac{E_x}{E_s} = \frac{I_x}{I_s} = \epsilon^{-\Gamma x} = \epsilon^{-Ax}\epsilon^{-jBx} = \epsilon^{-Ax}(\cos Bx - j \sin Bx) \tag{121}$$

The ratio of voltage to current at any point x along the line is

$$\frac{E_x}{I_x} = Z_0 \tag{122}$$

In the communication field, transmission lines may be classified according to the frequency bands that they are required to transmit, as *audio-* or *radio-frequency* lines. Simplified forms of the general transmission line formulas result from the introduction of approximations that are appropriate to each case.

45. Terminating Conditions for No Reflections and Maximum Power Transfer. Conditions for maximum power transfer through a transmission line call for the source and load impedances to be the *conjugates*, respectively, of the line input impedances as seen from each end of the line with the opposite termination connected. Conditions for no reflections at the ends of the line call for source and load impedances to be *equal* to the line characteristic impedance. These conditions are mutually incompatible if the line Z_0 is not pure resistance. In the general case where Z_0 has a reactance component, the best terminating impedances are a compromise between efficient energy transfer and the avoidance of reflections. Fortunately, in practical r-f lines Z_0 is very nearly a pure resistance so that both conditions are satisfied by making the source and load impedances pure resistances each equal to Z_0.

Radio-frequency lines whose terminal impedances are not equal to Z_0 are inefficient for two reasons: (1) they do not deliver maximum power to the load, and (2) some energy is reflected at each mismatched junction. Any impedance irregularity along the line also produces reflections and should usually be avoided. Thus, bends in an open-wire line should be gradual, with a minimum radius of about $\lambda/4$, and the line should be kept clear by at least $\lambda/4$ from large masses of conducting or dielectric materials, including ground.

The amount of reflection produced by a terminating impedance Z_T at the receiving end of a line of characteristic impedance Z_0 is expressed by the *reflection coefficient*

$$k = \frac{Z_T - Z_0}{Z_T + Z_0} = \frac{(Z_T/Z_0) - 1}{(Z_T/Z_0) + 1} \tag{123}$$

In lines whose principal purpose is conveyance of energy from one terminal to the other, reflections are objectionable because of increased ohmic and radiation losses. Also, the current and voltage peaks associated with the resulting standing waves must be considered in relation to conductor size and line insulation. The magnitude of the standing waves is given by the *standing-wave ratio* (SWR),

$$\text{SWR} = \frac{E_{\max}}{E_{\min}} = \frac{I_{\max}}{I_{\min}} = \frac{1 + |k|}{1 - |k|} \tag{124}$$

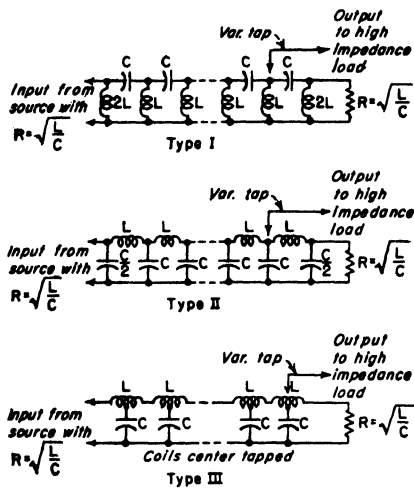
where $|k|$ is the positive value of the magnitude of the reflection coefficient as defined in Eq. (123), and the Es and Is refer to maximum and minimum voltages and currents on the line.

If the line attenuation is large, Eq. (124) holds true only at points near the receiving end, since the SWR will vary along the line length. In a low-loss line, however,

the SWR is essentially constant throughout; so the receiving end reflection coefficient may be determined by SWR measurements at the sending end, with reasonable accuracy.

Various methods for measuring SWR in transmission lines are in use. For a two-wire open line, voltage or current measurements at the nodes and antinodes give the necessary data. For coaxial lines, slotted sections of line with calibrated sliding probes give accurate results. For continuous observation of standing-wave conditions, reflectometers[1] are most convenient. Various forms of this device are available commercially, and they usually include a calibrated indicator for power in the line.

Fig. 72. Delay networks.

46. Delay Networks. Three typical networks are shown in Fig. 72. It is assumed that each is fed from a source whose internal impedance is $R = \sqrt{L/C}$ and that it is terminated in a pure resistance of the same value. The delayed signal output is taken from a variable tap and fed to any device having an internal impedance which is large compared to R. Since networks I and II are respectively high- and low-pass filters of the constant-K type, they have amplitude attenuation characteristics as shown in Sec. 65. Network III has no cutoff frequency if coupling between the two halves of each coil is perfect. Relative delay vs. frequency characteristics of the three types of networks (using the same values of L and C in each case) are shown in Fig. 73. The following relations are based on resistanceless networks and hold within about 1 per cent if the coil Q is 20 or more and the capacitor Q is 100 or more.

$$R = \sqrt{\frac{L}{C}} \quad \text{(approx) for all three networks} \tag{125}$$

Network I. Amplitude attenuation is large up to

$$f_c = \frac{1}{4\pi \sqrt{LC}} \quad \text{cycles} \tag{126}$$

and is zero thereafter (high-pass).

Delay for $f \leqq f_c$ is

$$t = \frac{1}{2f} \quad \text{sec/section} \tag{127}$$

Delay for $f > f_c$ is

$$t = \frac{2}{\omega} \sin^{-1} \frac{1}{2\omega \sqrt{LC}} \quad \text{sec/section exactly (angle in radians)} \tag{127a}$$

$$= \frac{1}{\omega^2 \sqrt{LC}} + \frac{1}{\omega}\left[0.04167 \left(\frac{1}{\omega \sqrt{LC}}\right)^3 + 0.00468 \left(\frac{1}{\omega \sqrt{LC}}\right)^5 + 0.000698 \left(\frac{1}{\omega \sqrt{LC}}\right)^7\right] \quad \text{sec/section (approx)} \tag{127b}$$

[1] Korman, Nathaniel I., Note on a Reflection-coefficient Meter, *Proc. IRE*, September, 1946, p. 657.

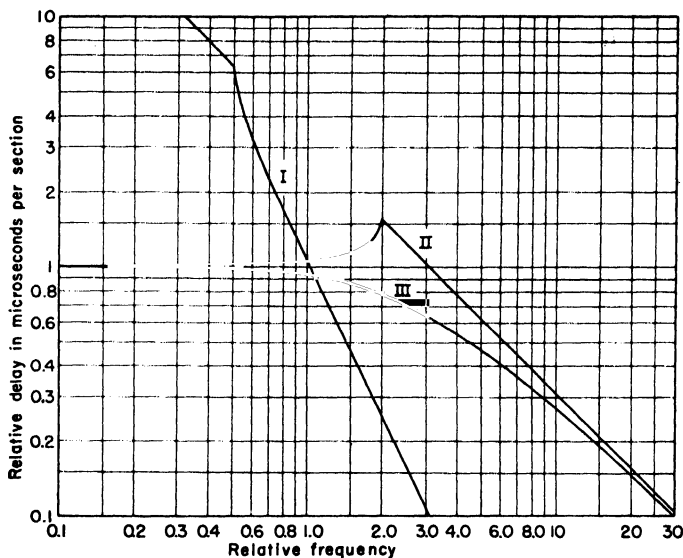



FIG. 73. Delay vs. frequency characteristics of three types of delay networks using the same LC in each.

Delay for $f \gg f_c$ is

$$t = \frac{1}{\omega^2 \sqrt{LC}} \quad \text{sec/section (approx)} \tag{127c}$$

Network II. Amplitude attenuation is zero up to cutoff frequency

$$f_c = \frac{1}{\pi \sqrt{LC}} \quad \text{cycles} \tag{128}$$

and is large thereafter (low-pass).

Delay for $f < f_c$ is

$$t = \frac{2}{\omega} \sin^{-1} \frac{\omega}{2} \sqrt{LC} \quad \text{sec/section exactly (angle in radians)} \tag{128a}$$

$$= \sqrt{LC} + \frac{1}{\omega} [0.04167(\omega \sqrt{LC})^3 + 0.00468(\omega \sqrt{LC})^5 + 0.000698(\omega \sqrt{LC})^7]$$

sec/section (approx)

Delay for $f \ll f_c$ is

$$t = \sqrt{LC} \quad \text{sec/section (approx)} \tag{128b}$$

Delay for $f \geqq f_c$ is

$$t = \frac{1}{2f} \quad \text{sec/section (approx)} \tag{128c}$$

Network III. Amplitude attenuation is zero at all frequencies for 100 per cent coupling between the two halves of each coil. In practice, this condition cannot be attained, and the attenuation will increase gradually with frequency.

Delay for all frequencies is

$$t = \frac{1}{\omega} \cos^{-1} \frac{4 - \omega^2 LC}{4 + \omega^2 LC} \quad \text{sec/section (angle in radians)} \tag{129}$$

By reference to Fig. 73 it is evident that the delay for networks II and III is constant to within ±5 per cent for all frequencies up to that indicated as relative frequency = 1. This value corresponds to $f = 1/(2\pi\sqrt{LC})$ for the network characteristics shown in the figure.

The principal use of network I is to obtain large delays at a single frequency, since in the vicinity of $f = 0.55/(2\pi\sqrt{LC})$ (relative frequency = 0.55 in Fig. 73) it produces about four times as much delay per section as II or III with the same coils and capacitors.

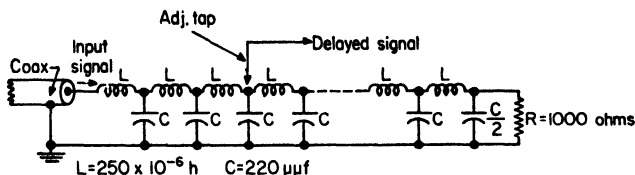



Fig. 74. Example of low-pass delay network used in television pulse circuits.

A practical application of a delay network of type II, as used in the formation of television synchronizing pulses, is illustrated in Fig. 74. With the constants shown the delay is 0.235 μsec per section for all frequencies up to about 680 kc.

47. *RC* Filter for Small Currents. An economical *RC* filter for small currents as suggested by Scott[1] is shown in Fig. 75. A special feature is the shunting circuit through R_A to feed voltage 180 deg out of phase to the point X. This can be adjusted to give a very high attenuation at one particular frequency which it is desired to eliminate. As shown, this filter is low-pass; a similar high-pass structure can be made by transposing the *R*'s and *C*'s.

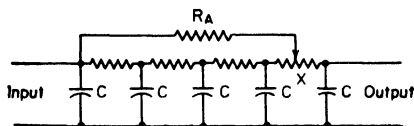



Fig. 75. *RC* filter for small currents.

48. Resistance Pads. Resistance pads are artificial lines whose series and shunt elements are pure resistances and are used principally as attenuators in a-f circuits. The amount of loss caused by insertion of a pad in a circuit may be accurately computed and is independent of frequency if the terminating impedances are resistances.

Either π or T structures may be used as pads, as shown in Fig. 76*a*. Both are electrically equivalent, but for identical values of loss and impedance one type may require resistors of more convenient values than the other. A pad to be used in a circuit that is balanced to ground should be of the balanced π or T type; otherwise the unbalanced network is satisfactory and requires several less resistors to build.

49. Pad Design. To design a pad, three constants must be known: the input and output impedances and the loss in decibels. The input and output impedances of a pad are usually made equal to those of the circuit to be connected to it. The design procedure depends upon whether these are equal or are different from each other.

1. *Equal Input and Output Impedances.* In this case, the value of each element is found by multiplying the proper constants, selected from Table 1 in connection with Fig. 76*a*, by the value of the input or output impedance Z in ohms.

Example: To design a 10-db, 500/500-ohm pad of the balanced T type: From Table 1, for 10-db attenuation, $a = 0.5195$ (hence $a/2 = 0.2597$) and $b = 0.7027$. Then the required resistances are $0.2597 \times 500 = 129.85$ for the series elements and $0.7027 \times 500 = 351.35$ ohms for the shunt element.

2. *Unequal Input and Output Impedances.* In this case, the design involves more computation. The value of each element is indicated by Fig. 76*b*, the constants of which are to be found in Table 1. The ratio of input to output impedance (or vice versa) of a pad of given loss is limited by the fact that, for large values of the impedance

[1] Scott, H. H., *Electronics*, August, 1939.

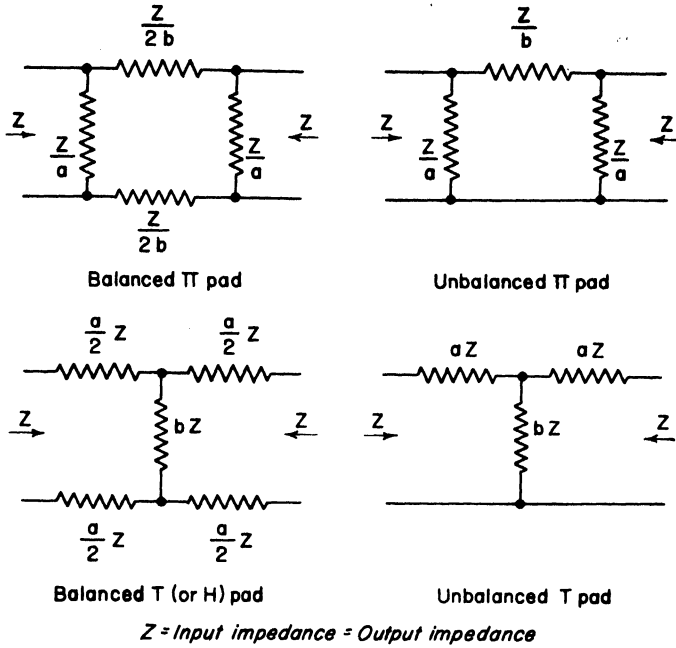



FIG. 76*a*. Pads to be used between equal impedances.

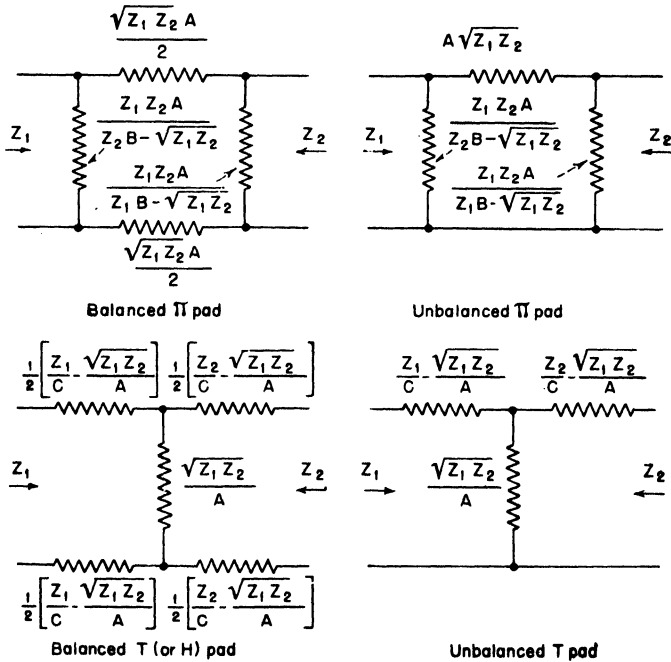



FIG. 76*b*. Pads to be used between unequal impedances.

ratio, certain of the pad resistors would have to be negative in value if the loss of the pad were to be below a certain minimum value. The maximum impedance ratio which a 10-db pad can have, for example, is 3.018. Stated in another way, this means that, if the impedance ratio of a pad is to be 3.018, its loss must be at least 10 db. The maximum impedance ratios for various values of pad losses are also given in Table 1. These are the same for both π and T pads.

Table 1. Constants for Pads of Fig. 76

Loss, db	A	B	C	a	b	$1/b$	$1/a$	$1/2b$	Max ratio Z_1/Z_2 or Z_2/Z_1
1	0.1154	1.007	0.1150	0.0575	8.664	0.1154	17.39	0.0577	1.014
2	0.2323	1.027	0.2263	0.1146	4.305	0.2323	8.724	0.1161	1.055
3	0.3523	1.060	0.3325	0.1710	2.838	0.3532	5.848	0.1761	1.124
4	0.4770	1.108	0.4305	0.2263	2.097	0.4770	4.419	0.2385	1.228
5	0.6084	1.170	0.5192	0.2801	1.645	0.6084	3.570	0.3042	1.369
6	0.7472	1.248	0.5986	0.3323	1.339	0.7472	3.009	0.3736	1.557
7	0.8960	1.343	0.6673	0.3825	1.116	0.8960	2.615	0.4480	1.804
8	1.0570	1.455	0.7264	0.4305	0.9462	1.0570	2.323	0.5285	2.117
9	1.2320	1.586	0.7763	0.4762	0.8118	1.2320	2.100	0.6160	2.515
10	1.4218	1.738	0.8181	0.5195	0.7027	1.4218	1.925	0.7109	3.018
11	1.6324	1.914	0.8527	0.5601	0.6127	1.6324	1.785	0.8162	3.663
12	1.8659	2.117	0.8814	0.5986	0.5359	1.8659	1.670	0.9329	4.482
13	2.1223	2.346	0.9046	0.6343	0.4712	2.1223	1.576	1.0611	5.504
14	2.4067	2.605	0.9235	0.6672	0.4155	2.4067	1.498	1.2033	6.786
15	2.7230	2.901	0.9387	0.6981	0.3672	2.7230	1.432	1.3615	8.415
20	4.9522	5.052	0.9802	0.8182	0.2020	4.9522	1.222	2.4761	25.52
25	8.8612	8.918	0.9940	0.8932	0.1128	8.8612	1.119	4.4306	79.52
30	15.800	15.830	0.9980	0.9387	0.06331	15.800	1.065	7.900	250.5
35	28.094	28.112	0.9994	0.9649	0.03560	28.094	1.036	14.047	790.2
40	50.000	50.0094	0.9998	0.9802	0.020000	50.000	1.020	25.000	2,500
45	88.928	88.933	0.9999	0.9888	0.01124	88.928	1.011	44.464	7,909
50	158.1	158.102	1.0000	0.9937	0.006325	158.10	1.006	79.050	24,980
60	500	500	1.0000	0.9980	0.002000	500	1.002	250	
70	1,581	1,581	1.0000	0.9994	0.000632	1,581	1.001	790	
80	5,000	5,000	1.0000	0.9998	0.000200	5,000	1.000	2,500	
90	15,810	15,810	1.0000	0.9999	0.0000632	15,810	1.000	7,905	
100	50,000	50,000	1.0000	1.0000	0.0000200	50,000	1.000	25,000	

$$A = \sinh \theta \qquad a = \frac{1}{C} - \frac{1}{A}$$

$$B = \cosh \theta$$

$$C = \tanh \theta \qquad b = \frac{1}{A}$$

$$\theta = \frac{\text{loss in db}}{8.686} \qquad \text{Max ratio } \frac{Z_1}{Z_2} \text{ or } \frac{Z_2}{Z_1} = B^2$$

Example: To design a 20-db 500/200-ohm pad of the unbalanced π type:

$$Z_1 = 500 \text{ ohms} \qquad Z_2 = 200 \text{ ohms}$$

From Table 1, $A = 4.9522$ and $B = 5.0522$. Then,

$$\text{Input shunt element} = \frac{Z_1 Z_2 A}{Z_2 B - \sqrt{Z_1 Z_2}} = 713 \text{ ohms}$$

$$\text{Series element} = \sqrt{Z_1 Z_2} A = 1{,}567 \text{ ohms}$$

$$\text{Output shunt element} = \frac{Z_1 Z_2 A}{Z_1 B - \sqrt{Z_1 Z_2}} = 430 \text{ ohms}$$

50. General Properties of R-f Lines. Useful working formulas for r-f lines are obtained by making use of the fact that at radio frequencies ωL and ωC become very large in comparison with R and G, respectively. The results are approximations whose accuracy depends upon whether R and G are treated as zero or as small quantities.

Symbols used in r-f line relations:

L = inductance, henrys/unit length
C = capacitance, farads/unit length
R = series resistance, ohms/unit length
G = shunt conductance, mhos/unit length

NOTE: L, C, R, and G are values for one unit length of two conductors.

V = actual velocity of propagation
V.F. = velocity factor, actual velocity as fraction or per cent of speed of light
v = speed of light, 3×10^8 meters/sec
ω = $2\pi \times$ frequency, cycles
f = frequency, cycles
f_{Mc} = frequency, megacycles
d_o = inner diam of outer conductor } both in same units
d_i = outer diam of inner conductor } both in same units
d = diam of conductor } both in same units
s = spacing between conductors, center to center } both in same units
l = length of line in unit lengths
Unit length = arbitrary small length of line, as 1 cm or 1 ft
in. = (as subscript) dimensions in inches

1. *Dissipationless R-f Lines.* ($R = 0$ and $G = 0$; approximate conditions for air-insulated copper lines.)

Characteristic impedance is

$$Z_0 = \sqrt{\frac{L}{C}} = \frac{1}{vC} \quad \text{ohms} \tag{130}$$

Velocity of propagation is

$$V = \frac{1}{\sqrt{LC}} \quad \text{unit lengths/sec}$$
$$= v = 3 \times 10^8 \text{ meters/sec} \tag{131}$$

Propagation constant is

$$P = jB = j\omega\sqrt{LC} \tag{132}$$

Wavelength constant is

$$B = \omega\sqrt{LC} = \omega C Z_0 \quad \text{radians/unit length}$$
$$= \frac{\omega}{v} = 2.0944 \times 10^{-8} f \quad \text{radians/meter} \tag{133}$$

Wavelength (physical distance) is

$$\lambda = \frac{2\pi}{\omega\sqrt{LC}} = \frac{1}{f\sqrt{LC}} \quad \text{unit lengths}$$
$$= \frac{v}{f} = \frac{3 \times 10^8}{f} = \frac{300}{f_{Mc}} \quad \text{meters} \tag{134}$$

Retardation time or *delay* is

$$t = \frac{B}{\omega} = \sqrt{LC} \qquad \text{sec/unit length}$$

$$= \frac{1}{v} = 0.333 \times 10^{-8} \qquad \text{sec/meter} \tag{135}$$

2. *R-f Lines with Small Attenuation.* ($R \neq 0$ and $G \neq 0$; approximate conditions for lines with dielectrics other than air.)

Attenuation is

$$A = 4.343\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) = 4.343\left(\frac{R}{Z_0} + GZ_0\right) \qquad \text{db/unit length} \tag{136}$$

The other constants for lines with small attenuation are

$$Z_0 = \sqrt{\frac{L}{C}}\left[1 + j\left(\frac{G}{2\omega C} - \frac{R}{2\omega L}\right)\right] \qquad \text{ohms} \tag{137}$$

$$V = \frac{\text{V.F.}}{\sqrt{LC}} \qquad \text{unit lengths/sec}$$

$$= v \times \text{V.F.} \qquad \text{meters/sec} \tag{138}$$

$$B = \frac{\omega\sqrt{LC}}{\text{V.F.}} \qquad \text{radians/unit length}$$

$$= \frac{2.09 \times 10^{-8} f}{\text{V.F.}} \qquad \text{radians/meter} \tag{139}$$

$$\lambda = \frac{\text{V.F.}}{f\sqrt{LC}} \qquad \text{unit lengths}$$

$$= \frac{3 \times 10^8}{f} \times \text{V.F.} = \frac{300}{f_{\text{Mc}}} \times \text{V.F.} \qquad \text{meters} \tag{140}$$

$$t = \frac{\sqrt{LC}}{\text{V.F.}} \qquad \text{sec/unit length}$$

$$= \frac{0.333 \times 10^{-8}}{\text{V.F.}} \qquad \text{sec/meter} \tag{141}$$

Although it is possible to calculate the velocity factor (V.F.) approximately for a given line, usually it is more practical to use the actually measured value. See Sec. 57 and Table 2.

The foregoing relations are equally applicable to coaxial and parallel wire lines. For lines of specified configurations, L, C, and R can be evaluated in terms of physical dimensions to give the relations that follow. G is not readily computed and, if it is not negligible, should preferably be measured.

3. *Parallel Two-wire Open R-f Line.* (Neither side grounded; copper conductors; R and G small.)

$$Z_0 = 277 \log_{10} \frac{2s}{d} \qquad \text{ohms} \tag{142}$$

This holds if $s \geqq 10d$ and the height above ground is $10s$ or more. Values of Z_0 for various conductor sizes and spacings are shown in Fig. 77.

$$A = 0.867 \frac{\sqrt{f_{\text{Mc}}}}{d_{\text{in.}} \times Z_0} \qquad \text{db/100 ft} \tag{143}$$

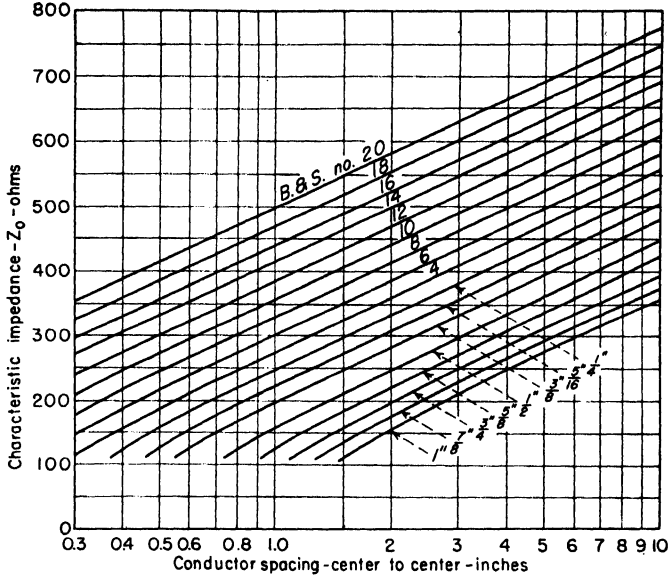



FIG. 77. Z_0 for balanced parallel copper-conductor r-f lines.

where $d_{in.}$ is the conductor diameter in inches. Equation (143) is based solely on copper losses and neglects all other losses. For accurate results, measured values are preferable to calculated values.

Storage factor is

$$Q = \frac{\omega L}{R} = 883.9 d_{in.} \times \sqrt{f_{Mc}} \times \log_{10} \frac{2s}{d} \tag{144}$$

Maximum voltage gradient is the same for both wires, occurs at the surface of each wire and at the point nearest the opposite wire, and is

$$\text{grad}_{max} = \frac{0.342E}{d_{in.} \times \log_{10} (2s/d)} \quad \text{volts/cm} \tag{145}$$

where E is the voltage between the wires at the point in question.

4. *Coaxial Two-conductor R-f Line.* (Copper conductors; R and G small.)

$$Z_0 = 138.5 \log_{10} \frac{d_o}{d_i} \quad \text{ohms} \tag{146}$$

Values of Z_0 vs. the ratio d_0/d_i are shown in Fig. 78.

$$A = \frac{0.00313 \sqrt{f_{Mc}} [1 + (d_o/d_i)]}{d_{o\ in.} \log_{10} (d_o/d_i)} \quad \text{db/100 ft} \tag{147}$$

where $d_{o\ in.}$ is the diameter of the outer conductor in inches. Equation (147) is based solely on copper losses and neglects leakage conductance G. When G cannot be neglected, measurements are preferable to calculations.

$$Q = \frac{0.8863 d_{o\ in.} \sqrt{f_{Mc}} \log_{10} (d_o/d_i)}{1 + d_o/d_i} \tag{148}$$

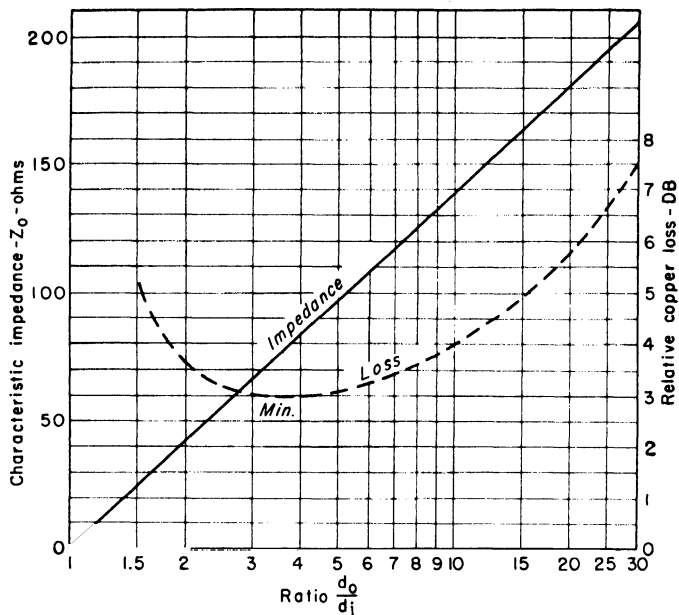



FIG. 78. Z_0 and relative loss vs. ratio of diameters for coaxial lines.

Optimum Diameter Ratios for Copper Coaxial Lines*

Optimum quantity	d_o/d_i	Z_0, ohms
Max antiresonant Z	9.185	132.9
Max Q and min attenuation (see Fig. 78)	3.592	76.64
Max breakdown volts	2.718	59.93
Min temp. rise, inner conductor	1.835	36.38
Max power carrying capacity	1.648	29.94

* Smith, P. H., *Electronics*, February, 1950.

Maximum voltage gradient occurs at the surface of the inner conductor and is

$$\text{grad}_{\text{max}} = \frac{0.342E}{d_{i\ \text{in.}} \log_{10} (d_o/d_i)} \quad \text{volts/cm} \tag{149}$$

where E is the voltage between inner and outer conductors and $d_{i\ \text{in.}}$ is the diameter of the inner conductor in inches.

5. *Resistance of Copper R-f Lines.* For two-wire open lines,

$$R = \frac{0.1996\ \sqrt{f_{\text{Mc}}}}{d_{\text{in.}}} \quad \text{ohms/100 ft} \tag{150}$$

For coaxial lines,

$$R = \frac{0.0998\ \sqrt{f_{\text{Mc}}}\ [1 + (d_o/d_i)]}{d_{o\ \text{in.}}} \quad \text{ohms/100 ft} \tag{151}$$

In both cases, the resistance is in terms of 100 ft of line and takes into account the fact that there are two conductors.

51. RCA Six-wire Line. A type of open-wire line frequently used for feeding a broadcast transmitting antenna which operates with one side grounded is shown in Fig. 80. The four outer wires are connected together and grounded to act as one conductor, and the two inner wires are connected together to act as the second conductor. The outer wires function as a shield and limit radiation from the line to about 0.015 per cent in one wavelength of line as compared to about 1 per cent for an unbalanced two-wire line of corresponding dimensions. Ground conductivity losses are also reduced by using the six-wire line. The characteristic impedance using No. 6 copper conductors is 230 ohms, and the attenuation is approximately 0.04 $\sqrt{f_{Mc}}$ db per 100 ft throughout the 550- to 1,600-kc broadcast band for a soil conductivity of 50×10^{-15} emu.

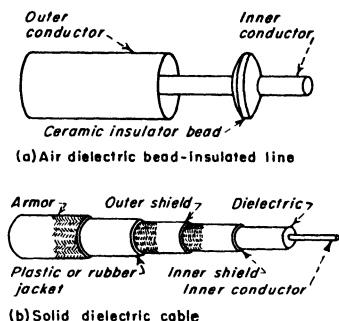



FIG. 79. Typical construction on commercial coaxial types of r-f transmission lines.

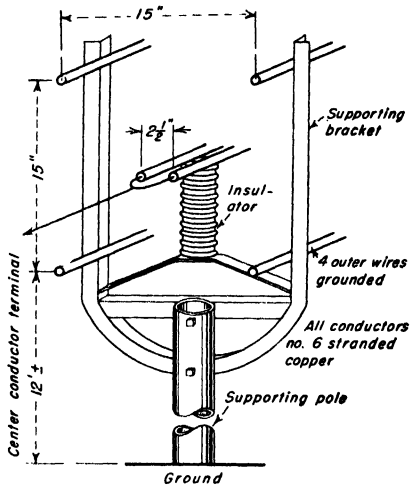



FIG. 80. Six-wire r-f line with low radiation. (*RCA*.)

52. Impedance Relations in R-f Lines. At high frequencies the attenuation constant of a line is so small that it may be neglected in many cases to simplify impedance calculations. On this basis, Z_0 is essentially pure resistance, and the input impedance of a line terminated at its end in any impedance Z_T reduces, from Eq. (117), to

$$\begin{aligned} Z_i &= Z_0 \frac{Z_T \cos lB + jZ_0 \sin lB}{Z_0 \cos lB + jZ_T \sin lB} \\ &= Z_0 \frac{Z_T \cos 2\pi l/\lambda + jZ_0 \sin 2\pi l/\lambda}{Z_0 \cos 2\pi l/\lambda + jZ_T \sin 2\pi l/\lambda} \\ &= Z_0 \frac{Z_T/Z_0 + j \tan 2\pi l/\lambda}{1 + jZ_T/Z_0 \tan 2\pi l/\lambda} \quad \text{ohms} \end{aligned} \tag{152}$$

in which l is the length of the line and λ is the wavelength of the signal, both measured in the same unit of length. This gives the value of $2\pi l/\lambda$ in radians. It is often convenient to substitute the electrical length of the line in degrees for $2\pi l/\lambda$, on the basis of 360 deg per wavelength. Thus a $\lambda/4$ line represents 90 electrical degrees, etc.

Actual numerical computation of input impedance by Eq. (152) is a tedious process, especially if Z_T is not pure resistance. When many such computations are to be made, the transmission line calculator of P. H. Smith[1] is a practical timesaver. This device is in the form of a circular slide rule and also yields such information as standing-wave ratios and line attenuation.

[1] Smith, P. H., Transmission Line Calculator, *Electronics*, January, 1939, and Improved Transmission Line Calculator, *Electronics*, January, 1944.

Table 2. Characteristics of Commercial R-f Transmission Lines
(Manufacturer's ratings)

Manufacturer	Type No. Mfr.	Type No. RG/—	Type line	Outside dimensions, in.	Insulation	Conductor size (inner, if coaxial)	Velocity*	Z_0	Max rms volts or power	Loss in decibels per 100 ft: 1 Mc	10 Mc	50 Mc	100 Mc	200 Mc	1,000 Mc	3,000 Mc
Amphenol	21-004	8	Coaxial	0.405 diam	Polyethylene	7—No. 21	65.9	52	4,000 volts	0.130	0.520	.....	2.10	.....	8.50	18.0
	21-007	11	Coaxial	0.405 diam	Polyethylene	7—No. 26	65.9	75	4,000 volts	0.115	0.470	.....	1.90	.....	7.70	16.0
	21-013	17	Coaxial	0.870 diam	Polyethylene	0.188 in.	65.9	52	11,000 volts	0.038	0.180	.....	0.85	.....	4.00	9.6
	21-015	19	Coaxial	1.12 diam	Polyethylene	0.250 in.	65.9	52	14,000 volts	0.030	0.142	.....	0.70	.....	3.40	8.3
	21-017	21	Coaxial	0.332 diam	Polyethylene	16 Nichrome	65.0	53	2,700 volts	1.3	4.3	.....	14.0	.....	46.0	82.0
	21-019	34	Coaxial	0.625 diam	Polyethylene	7—No. 21	65.9	72	5,200 volts	0.115	0.470	.....	1.90	.....	7.70	16.0
	21-024	58	Coaxial	0.195 diam	Polyethylene	No. 20	65.9	53.5	1,900 volts	0.240	1.00	.....	4.10	.....	17.0	34.0
	21-025	59	Coaxial	0.242 diam	Polyethylene	No. 22	65.9	73	2,300 volts	0.260	1.00	.....	3.75	.....	14.3	28.5
	21-038	22	Parallel shielded	0.405 diam	Polyethylene	7—0.0152 in. each	65.9	95	1,000 volts	0.220	0.860	.....	3.40			
RCA	6 wire	..	Open wire	15 square†	Ceramic post	No. 6 stranded	....	230	500 kw	0.04						
Andrew	83	..	Coaxial	0.375 diam	Spaced beads	No. 12	86.0	70	250 watts‡	0.100	0.324	0.754	1.09	1.61		
	87	..	Coaxial	1.625 diam	Spaced beads	0.50 in.	97.8	66	5 kw‡	0.020	0.065	0.150	0.219	0.321		
	737	..	Coaxial	0.875 diam	Spaced beads	0.25 in.	92.0	64	2 kw‡	0.037	0.119	0.281	0.414	0.617		
	450	..	Coaxial	0.875 diam	Spaced beads	0.312 in.	93.3	51.5	3 kw‡	0.041	0.130	0.318	0.460	0.673		
	451	..	Coaxial	1.625 diam	Spaced beads	0.625 in.	95.3	51.5	10 kw‡	0.021	0.068	0.161	0.234	0.346		
	452	..	Coaxial	3.125 diam	Spaced beads	1.2 in.	92.8	51.5	42 kw‡	0.011	0.035	0.096	0.145	0.230		
	453	..	Coaxial	6.125 diam	Spaced beads	2.5 in.	99.3	51.5	166 kw‡	0.0053	0.017	0.039	0.056	0.081		
Amphenol	14-056	..	Parallel conductor	0.4 × 0.062	Polyethylene	7—No. 28	82.0	300	Receiver type	......	0.390	1.30	2.10	3.60		
	14-079	..	Parallel conductor	0.185 × 0.06	Polyethylene	7—No. 28	77.0	150	Receiver type	......	0.460	1.60	2.70	4.70		
	14-080	..	Parallel conductor	0.133 × 0.06	Polyethylene	7—No. 28	68.0	75	Receiver type	......	0.850	3.00	5.00	8.30		

* Velocity in per cent free space velocity.
† Not including mounting brackets.
‡ For line operated with SWR = 1.
Data obtained through courtesy of American Phenolic Corp., Andrew Co., and RCA.

Open-circuited Line. The input impedance of a no-loss line of any length open-circuited at its far end ($Z_T = \infty$) is a pure reactance

$$Z_i = -jZ_0 \cot lB = -jZ_0 \cot 2\pi l/\lambda \tag{153}$$

Short-circuited Line. The input impedance of a no-loss line of any length short-circuited at its far end ($Z_T = 0$) is a pure reactance

$$Z_i = jZ_0 \tan lB = jZ_0 \tan 2\pi l/\lambda \tag{154}$$

The reactive input impedance of a short length of line, or "stub" line, provides a handy method of balancing out unwanted reactance in the termination of a h-f line.

(a) $Z_i = Z_0^2/Z_T$

(b) $Z_0 = \sqrt{Z_T Z_S}$

(c) Two or more sections for broad-band matching

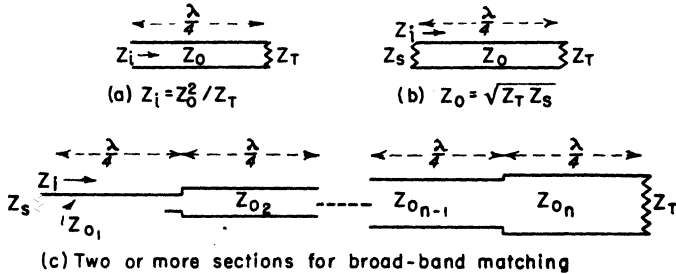



FIG. 81. Quarter-wave lines as impedance transformers. All sections are $\lambda/4$ at f_0.

The stub is connected in parallel across the terminating impedance and trimmed in length by trial until an increase in signal strength or similar indication shows that the desired result is obtained. Such stubs may be either open- or short-circuited at their free ends.

53. Properties of Quarter-wave R-f Lines. In dissipationless r-f lines the input impedance Z_i of a $\lambda/4$ line—and of all lines whose lengths are odd multiples of $\lambda/4$—is the reciprocal of its terminating impedance Z_T when the two impedances are expressed in terms of the line characteristic impedance Z_0 (Fig. 81). Thus

$$\frac{Z_i}{Z_0} = \frac{Z_0}{Z_T} \quad \text{or} \quad Z_i = \frac{Z_0^2}{Z_T} \tag{155}$$

This condition is independent of the value of the characteristic impedance Z_0 of the line.

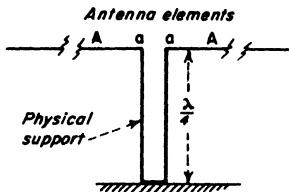



FIG. 82. Support for antenna elements A made into form of $\lambda/4$ line, effectively insulating points a-a from ground.

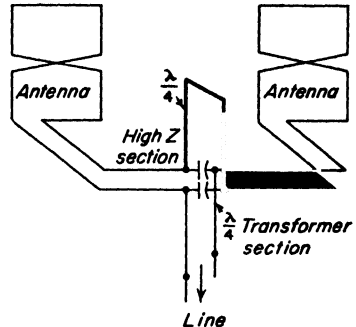



FIG. 83. Use of $\lambda/4$ short-circuited line to by-pass low-frequency currents for melting sleet without disturbing the r-f impedance of the system.

If a $\lambda/4$ line is short-circuited at its far end ($Z_T = 0$), Z_i is very large, approaching ∞ in the case of a dissipationless line. This fact is made use of in such practical applications as the "conductive insulator," Fig. 82, and the sleet-melting arrangement shown in Fig. 83.

If a $\lambda/4$ line is open-circuited at its far end ($Z_T = \infty$), Z_i is very small, approaching zero for a dissipationless line. A practical application is the third-harmonic suppressor shown in Fig. 84, in which an open-circuited $\lambda/4$ line A (at third-harmonic frequency) is shunted across a line carrying a signal of fundamental frequency f. The stub line effectively short-circuits the main line for currents of frequency $3f$, thereby eliminating the third harmonic. However, the effect of the stub at the fundamental frequency f must also be considered. For this frequency, the stub length is $\lambda/12$, and the line presents a shunt reactance of $-j1.732Z_0$ ohms to currents of frequency f. To balance out this reactance a second stub B, of length $\lambda/6$ and short-circuited at its far end, is also shunted across the main line at the same point. This presents an equal and opposite reactance of $+j1.732Z_0$ ohms to resonate at frequency f with the reactance of stub A. Stub B has the length $\lambda/2$ at the harmonic frequency and, therefore, acts merely as an additional short circuit across the line at this frequency. An open-circuited line of length $5\lambda/12$ might also be used for stub B.

FIG. 84. Use of stub transmission line to suppress third harmonic.

54. Quarter-wave R-f Lines as Impedance Transformers. A $\lambda/4$ section of line can be used as an impedance-matching or transforming device by virtue of its impedance-inversion properties (Sec. 53). To match an impedance Z_T to another Z_S, a $\lambda/4$ section of transmission line is inserted between the two impedances (Fig. 81b). The characteristic impedance of the matching stub is determined by

$$Z_0 = \sqrt{Z_T Z_S} \tag{156}$$

A perfect transformation is obtained at only the one frequency where the line is exactly $\lambda/4$, and at odd integral multiples of this frequency.

55. Band-pass Characteristics of λ/4 R-f Line Transformers. A line of given physical length has an electrical length of $\lambda/4$ at only one frequency f_0. (Odd integral multiples of f_0 giving odd multiples of $\lambda/4$ have the same transformation properties but are ignored here for practical purposes.) In a system employing a $\lambda/4$ line as an impedance transformer, the impedance match will be less nearly perfect as the frequency deviates from f_0. This point is of interest in connection with modulated signals and band-pass systems. The resulting impedance mismatch at frequencies above and below f_0 is represented for a single $\lambda/4$ section by curve 1 of Fig. 85, which shows the mismatch in terms of magnitude and phase angle for a 10:1 impedance transformation. Frequency is indicated by the ratio f/f_0, and the degree of matching by Z_i/Z_s, where Z_i is the transformed value of Z_T as seen from the input end of the line, and Z_s is the impedance to which Z_T is to be matched.

The degree of matching can be improved over a band of frequencies by using two or more $\lambda/4$ sections (*i.e.*, $\lambda/4$ at f_0) in succession to effect the transformation as shown in Fig. 81. The Z_0 of the successive sections must be selected according to one of several possible schemes. The curves in Fig. 85 show the matching conditions obtained by the use of from two to six $\lambda/4$ sections for an impedance transformation ratio of 10:1 when the successive Z_0's follow an exponential taper law in which the Z_0 of each section is found—in terms of the sending end impedance Z_s—from the relation

$$\log_{10} \frac{Z_{0_n}}{Z_s} = \frac{(2n - 1) \log_{10} M}{2n} = a \log_{10} M \tag{157}$$

where n = number of $\lambda/4$ section counted from sending end
M = over-all transformation ratio, Z_T/Z_s
a = constant found from Table 3

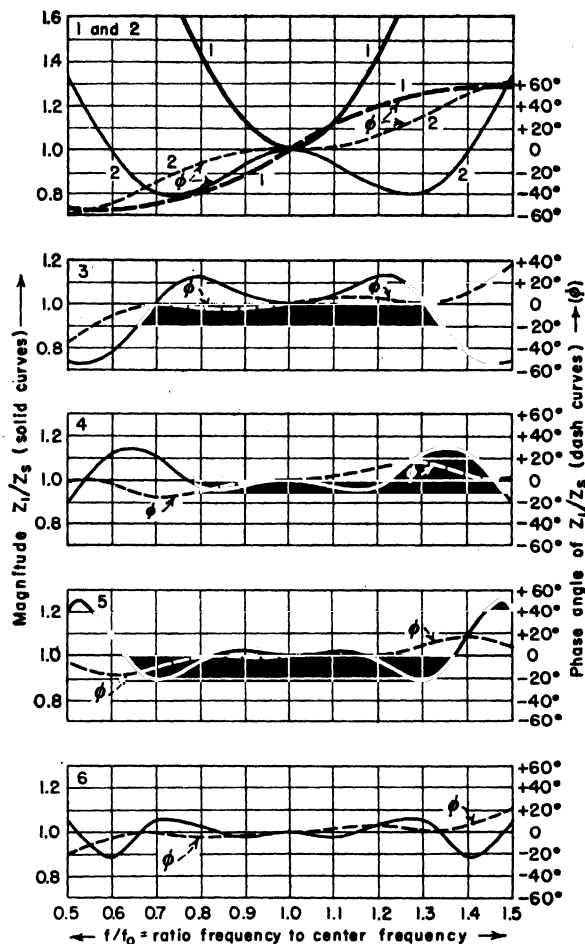



FIG. 85. Degree of wide-band impedance match obtained by use of from one to six $\lambda/4$ transformer sections for 10:1 impedance ratio.

Table 3. Values of a for Use in Eq. (157)

Values of a	Number of sections					
	1	2	3	4	5	6
For Z_{0_1}/Z_s	0.5	0.25	0.1666	0.125	0.1	0.0833
For Z_{0_2}/Z_s	...	0.75	0.5	0.375	0.3	0.25
For Z_{0_3}/Z_s	...	....	0.8333	0.625	0.5	0.4166
For Z_{0_4}/Z_s	...	....	......	0.875	0.7	0.5833
For Z_{0_5}/Z_s	...	....	......	.....	0.9	0.75
For Z_{0_6}/Z_s	...	....	......	.....	...	0.9166

In Fig. 85 the number attached to each curve represents the total number of $\lambda/4$ sections employed. The curves shown are for a transformation ratio of 10:1. For a smaller ratio, they will be flatter; for a larger ratio, they will show larger variations within a given range of frequency. In the design of an exponentially tapered $\lambda/4$ line transformer to match an impedance Z_T to another impedance Z_s, it is convenient to use the last expression in Eq. (157) together with Table 3, which gives values of a for transformers of from one to six sections. The value of f_0 should be chosen in the center of the frequency band to be covered by the system, and the physical length of each section of line should correspond to an electrical length of $\lambda/4$ at the frequency f_0.

56. Properties of Half-wave R-f Lines. Dissipationless half-wave lines and lines whose lengths are integral multiples of $\lambda/2$ have input impedance Z_i exactly equal to the terminating impedance Z_T, independent of the Z_0 of the line.

$$Z_i = Z_T \qquad \text{ohms} \tag{158}$$

57. Dimensional Data for R-f Stub Lines. The range of frequencies in which the use of stub or transformer lines is practical is limited at the lower extreme by the physical dimensions of the line structure required at the longer wavelengths, and at the upper extreme by the relative inefficiency of conductor-type lines as compared to wave guides.

To determine the physical length of a stub line, it is necessary to take into account the actual velocity of propagation along the line. The following relations are useful in determining physical lengths of lines for this purpose:

$$\lambda/4 \text{ length} = \frac{75}{f_{Mc}} \times \text{V.F. meters} = \frac{246.1}{f_{Mc}} \times \text{V.F. ft} = \frac{2{,}952.8}{f_{Mc}} \times \text{V.F. in.}$$

$$\lambda/2 \text{ length} = \frac{150}{f_{Mc}} \times \text{V.F. meters} = \frac{492.1}{f_{Mc}} \times \text{V.F. ft} = \frac{5{,}905.5}{f_{Mc}} \times \text{V.F. in.}$$

$$\lambda \text{ length} = \frac{300}{f_{Mc}} \times \text{V.F. meters} = \frac{984.3}{f_{Mc}} \times \text{V.F. ft} = \frac{11{,}811}{f_{Mc}} \times \text{V.F. in.}$$

where f_{Mc} = frequency, Mc
λ = 1 electrical wavelength
V.F. = velocity factor expressed as a fraction of the speed of light

The value of V.F. must be determined for the line to be used. Representative values for common types of lines are

V.F. = 0.92 to 0.99 for open-wire air dielectric lines
V.F. = 0.79 to 0.99 for beaded coaxial lines with air dielectric
V.F. = 0.6 to 0.82 for solid dielectric lines

58. Voltage and Power Ratings of R-f Lines. In selecting a transmission line for a transmitter application, consideration must be given to voltage flashover and power limitations. Maximum voltage is limited by conductor spacing, insulation and the corona formation point, and maximum power by line losses and conductor current-carrying capacity. Voltage, current, and power are interrelated in the case of a line operated with matched terminal impedances so that standing waves do not exist:

$$P = \frac{E^2}{Z_0} = I^2 Z_0 \qquad \text{watts} \tag{159}$$

where P = power carried by line, watts
E = rms voltage between conductors, volts
I = rms current in each conductor, amp
Z_0 = characteristic impedance of line, ohms

The corresponding relations for a line with standing waves are

$$E_{max} = \sqrt{P \times Z_0 \times \text{SWR}} \tag{160}$$

$$I_{max} = \sqrt{\frac{P \times \text{SWR}}{Z_0}} \tag{161}$$

where SWR is the standing-wave ratio as defined in Eq. (124).

Corona occurs when the voltage gradient at some point in the line exceeds the breakdown potential of air (or gas in the case of gas-filled lines). Corona is not necessarily accompanied by flashover but is objectionable because the energy dissipated represents an increase in the line losses. Chemical by-products of sustained corona discharges may also tend to deteriorate the line insulation. The critical surface gradient at which corona begins to form in air at 25°C and atmospheric pressure depends upon the conductor size. For conductors of No. 0 B&S and larger, it is about 40,000 volts per cm peak, or 28,000 volts per cm rms. For smaller conductors, it is approximately $(40{,}000 + N \times 1{,}670)$ volts per cm peak, or $(28{,}000 + N \times 1{,}180)$ volts per cm rms, where N is the B&S gage size of the wire. Maximum gradients in r-f lines occur at the surfaces of the conductors in open-wire lines, and at the surface of the inner conductor in coaxial lines. They may be evaluated by Eqs. (145) and (149). All such computations must be considered as approximations, but they are sufficiently accurate for engineering purposes if adequate safety factors are allowed.

WAVE FILTERS

59. Wave filters are forms of artificial lines, such as those of Fig. 86, purposely designed to transmit efficiently current in a desired band of frequencies and more or less completely to suppress all other frequencies. The boundary frequencies between transmission bands and attenuation bands are called *cutoff frequencies*.

The following brief discussion of wave-filter design is intended to serve as a guide to the design of simple filters for use where the requirements are not very severe. For complete information concerning the design of filters to meet more exacting specifications, the references listed in the bibliography at the end of this chapter should be consulted.

FIG. 86. Basic wave-filter circuits.

Filters are divided into four classes, according to the frequency bands which they are intended to transmit: *low pass, high pass, band pass,* and *band elimination.*

60. Losses in Filters, and Effects of Dissipation. The elements of ideal wave filters are always pure reactances; practically, however, some dissipation must always be tolerated owing to the resistance of coils and capacitors, but this is made as small as possible by employing high-Q elements.

The terminating impedances of a filter are usually resistances equal in value to the image impedances of the filter. Then the loss within the transmitted bands (except near the cutoff frequency) is mainly due to dissipation in the elements and is usually small. In the vicinity of cutoff and the point of maximum attenuation, the total insertion loss of a filter involves the reflection and interaction losses as well as the attenuation. The loss elsewhere in the attenuated bands is very nearly the sum of the attenuation constants of the various sections, minus a gain of approximately 6 db which is due to reflections resulting from impedance mismatches occurring in these regions. Methods for the exact calculation of filter losses are beyond the scope of this handbook but are available in the published works of Zobel, Johnson, and Shea.

61. The Basic Filter Section. The basis of filter design is the full L section, consisting of a series element Z_1 and a shunt element Z_2 as shown at a in Fig. 87. The relation of such a section to an infinite line is also indicated. In a wave filter, where the

number of sections is finite and small instead of infinite, symmetrical sections are used. These are either T or π networks as shown at *b* and *c* in Fig. 87. The T section may be considered as being cut from the infinite line (Fig. 87*a*) at the mid-points (*a–a*) of two consecutive series elements Z_1 and is said to be "mid-series terminated." The π section may be considered as being cut at the midpoints (*b–b*) of two consecutive shunt elements and is said to be "mid-shunt terminated." (To form a mid-shunt termination, each full-shunt element is replaced by an equivalent two impedances

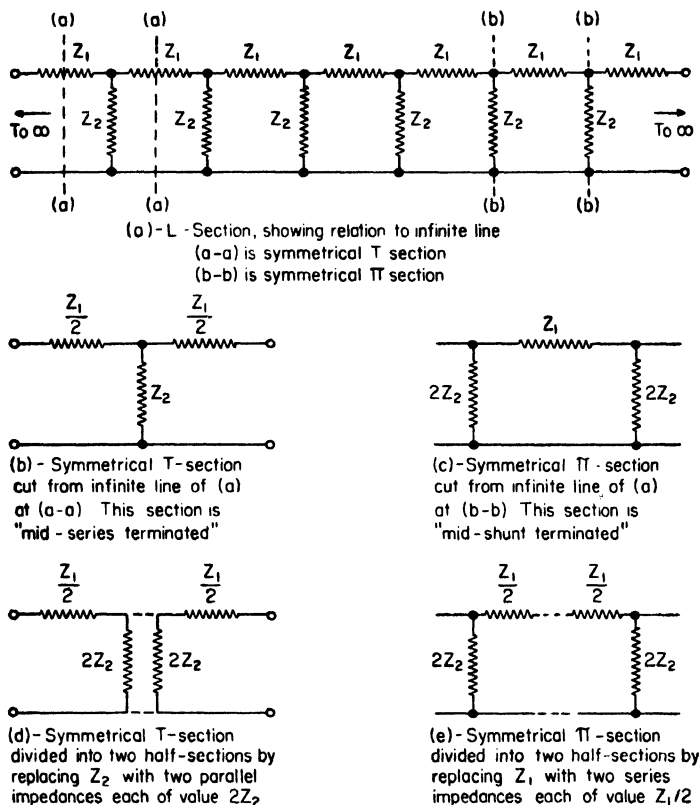



FIG. 87. Equivalence of T and π networks.

in parallel, each of value $2Z_2$.) Either a T or π section may be divided into pairs of equivalent half sections as shown at *d* and *e* in Fig. 87.

62. Types of Sections. 1. *Constant-K Sections.* The simplest and most common type of filter section is that in which the impedances Z_1 and Z_2 are so related that their product is a constant

$$Z_1 \times Z_2 = K^2$$

at all frequencies. From this it derives its name "constant-K" section. The configuration and circuit constants of the four classes of constant-K sections are shown in the filter-design formulas in Sec. 66. The image impedance of mid-series and mid-shunt terminated constant-K sections within the transmission bands are functions of frequency, but each approaches the value K at some frequency within the band. The value K is therefore taken as the nominal resistance of the constant-K section for design purposes. If a constant-K section is used with one or both of its terminals

connected to a pure resistance of value $R = K$, the impedances will be mismatched for all frequencies within the transmitted band except one, and the actual insertion or transmission loss of the filter will be increased by reflection losses at the terminations. This causes an even more gradual cutoff for the constant-K section than its attenuation curve would indicate.

2. *m-derived Sections.* In many filters, a sharper cutoff than that given by a constant-K type of structure is required. Such a characteristic may be realized in the so-called *m-derived section*, which is due to Otto J. Zobel.[1] This type of section is derived from the constant-K section as a prototype but is made to have sharper cutoff than the prototype by the addition of impedance elements in either the shunt or series arms so that infinite attenuation occurs at some frequency beyond cutoff. Each impedance of the m-type section is related to those of the constant-K section by a factor which is a function of a constant m. The latter is in turn a function of the ratio between the frequency of infinite attenuation and the cutoff frequency and may have any value between 0 and plus 1. The sharpness of cutoff increases as m approaches 0. This effect is illustrated in Fig. 88 for various values of m. It will be noted that, when m is equal to 1, the structure is identical with the constant-K structure. Also, from Fig. 88, it appears that from the viewpoint of obtaining a uniform degree of attenuation throughout the attenuated band the combination of a constant-K section ($m = 1$) (having gradual cutoff but large attenuation remote from cutoff) with one having a small value of m and sharp cutoff ($m = 0.3$, for example) would be desirable. This principle is valuable in the design of composite filters.

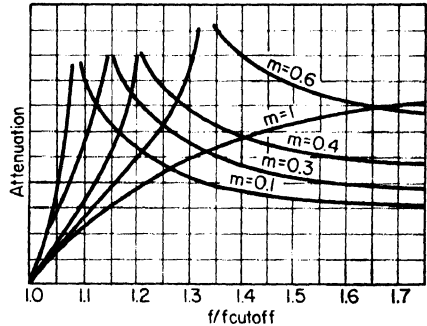



FIG. 88. Effect of m on sharpness of cutoff in a low-pass filter structure.

3. *Shunt-derived and Series-derived m Sections.* Two forms of m-derived sections exist; if the extra impedance is added to the shunt arm, the section is called *series derived*, while, if it is added to the series arm, the section is called *shunt derived*. (See illustrations of derived sections under Filter-design Formulas, Sec. 66.)

63. Assembly of Sections into Filters. A filter may consist of any number of sections from a single one-half section to five or six full sections, depending on the amount of attenuation of unwanted frequencies required. The *amount* of attenuation in the rejected band depends upon the number of filter sections used, and the *shape* of the transmission curve depends upon the types of sections employed.

64. One-half- and One-section Filters. If a half section or one full section is used alone as a filter and the requirements regarding the cutoff are not too sharp, an m-derived section is usually preferable, with $m = 0.6$. This will provide the best impedance match with resistance termination. Either of the structures shown in Fig. 89 is suitable for use with terminations of resistance R.

65. Multisection Filters. Filters having more than one section are of two types:

A *uniform* filter is one in which all sections are identical with the exception of the end sections. The latter are ordinarily half sections suitable for connecting the filter to its terminating resistances.

A *composite* filter is one made up of two or more sections having different characteristics, each of which is designed to contribute some special property to the characteristic of the filter as a whole. For example, one section which has sharp cutoff but a diminishing attenuation beyond cutoff may be combined with another section having a gradual cutoff and increasing attenuation beyond, as shown at I and II in Fig. 90. The resulting composite structure will then have both sharp cutoff and high attenuation beyond, as shown at III. In general, constant-K sections have gradual cutoffs

[1] *Bell System Tech. J.*, January, 1923.

FIG. 89. Half- and full-section filters.

with increasing attenuation beyond, while m sections with small values of m have the sharpest cutoff characteristics. Still other types of sections may be added to match impedances at the junctions of the filter and its terminating resistances, or to further alter the transmission characteristics.

In a composite filter it is essential that the image impedances be matched at each junction of the component sections, to avoid reflection losses which would impair the transmission curve of the filter. Likewise, the end terminations of the filter should as nearly as possible match the terminating resistances. One of the principal advantages of the m-type structure is that its image impedances can be made identical with other m-type sections or with constant-K sections; or they can be made to approximate resistances over the transmission band for terminating purposes. A complete analysis of the impedance conditions within a wave filter is not possible in the limited space available here but may be found in the references listed at the end of this chapter. The following will suffice as working rules in designing simple filters for ordinary requirements:

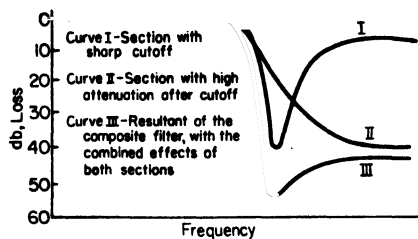
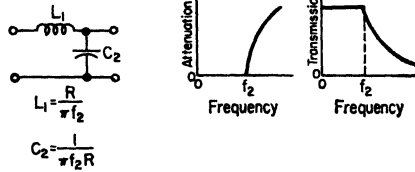
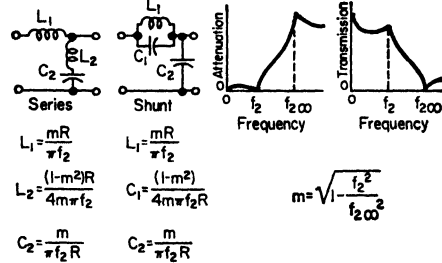
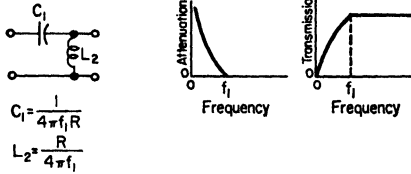
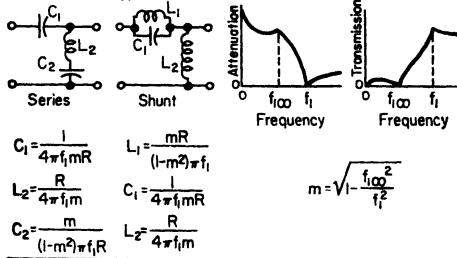
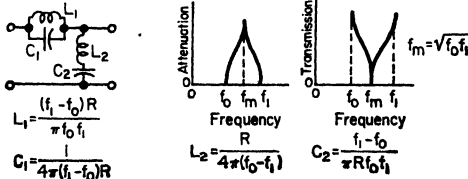



FIG. 90. Transmission curves for composite low-pass filter.

End Terminations. *Resistance.* A mid-shunt termination of a series-derived m-type section or half section, or a mid-series termination of a shunt-derived section or half section, with $m = 0.6$ in either case.

For Parallel or Series Connection with Other Filters. An 0.8 series constant-K section or half section (*i.e.*, one terminated in a series arm equal to 0.8 of a full series arm, Z_1).

Here, as well as in the two preceding paragraphs, the image impedance of the internal section next to the end section in either case must match the image impedance at the inner terminals of the latter, in accordance with the following.

I LOW PASS FILTERS
(a)-Constant-k type
L1
C2
L1 = R / πf2
C2 = 1 / πf2R
Attenuation
Transmission
0
f2
Frequency
(b) m-Derived type
L1
L2
C2
C1
Series
Shunt
L1 = mR / πf2
L2 = (1-m2)R / 4mπf2
C2 = m / πf2R
L1 = mR / πf2
C1 = (1-m2) / 4mπf2R
C2 = m / πf2R
f2∞
m = √(1 - f2^2 / f2∞^2)
II-HIGH PASS FILTERS
(a)-Constant-k type
C1 = 1 / 4πf1R
L2 = R / 4πf1
f1
(b)-m-Derived types
C1 = 1 / 4πf1mR
L2 = R / 4πf1m
C2 = m / (1-m2)πf1R
L1 = mR / (1-m2)πf1
C1 = 1 / 4πf1mR
L2 = R / 4πf1m
f1∞
m = √(1 - f1∞^2 / f1^2)
III-BAND ELIMINATION FILTERS
(a) Constant-k type
L1 = (f1 - f0)R / πf0f1
C1 = 1 / 4π(f1 - f0)R
L2 = R / 4π(f0 - f1)
C2 = f1 - f0 / πRf0f1
f0
fm
fm = √(f0f1)

III-BAND ELIMINATION FILTERS (continued)

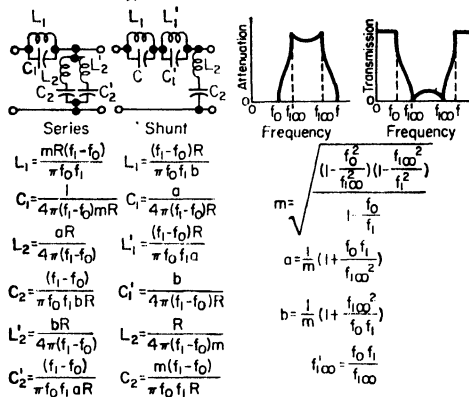
(b)-m-Derived types

Series	Shunt
$L_1 = \frac{mR(f_1 - f_0)}{\pi f_0 f_1}$	$L_1 = \frac{(f_1 - f_0)R}{\pi f_0 f_1 b}$
$C_1 = \frac{1}{4\pi(f_1 - f_0)mR}$	$C_1 = \frac{a}{4\pi(f_1 - f_0)R}$
$L_2 = \frac{aR}{4\pi(f_1 - f_0)}$	$L'_1 = \frac{(f_1 - f_0)R}{\pi f_0 f_1 a}$
$C_2 = \frac{(f_1 - f_0)}{\pi f_0 f_1 bR}$	$C'_1 = \frac{b}{4\pi(f_1 - f_0)R}$
$L'_2 = \frac{bR}{4\pi(f_1 - f_0)}$	$L_2 = \frac{R}{4\pi(f_1 - f_0)m}$
$C'_2 = \frac{(f_1 - f_0)}{\pi f_0 f_1 aR}$	$C_2 = \frac{m(f_1 - f_0)}{\pi f_0 f_1 R}$

$$m = \sqrt{\frac{\left(1 - \frac{f_0^2}{f_{1\infty}^2}\right)\left(1 - \frac{f_{1\infty}^2}{f_1^2}\right)}{1 - \frac{f_0}{f_1}}}$$

$$a = \frac{1}{m}\left(1 + \frac{f_0 f_1}{f_{1\infty}^2}\right)$$

$$b = \frac{1}{m}\left(1 + \frac{f_{1\infty}^2}{f_0 f_1}\right)$$

$$f'_{1\infty} = \frac{f_0 f_1}{f_{1\infty}}$$

IV BAND PASS FILTERS

(a)-Constant-k type

$$L_1 = \frac{R}{\pi(f_2 - f_1)} \qquad L_2 = \frac{(f_2 - f_1)R}{4\pi f_2 f_1}$$

$$C_1 = \frac{(f_2 - f_1)}{4\pi f_2 f_1 R} \qquad C_2 = \frac{1}{\pi(f_2 - f_1)R}$$

(b)-m-Derived types

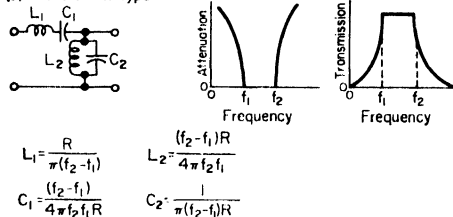
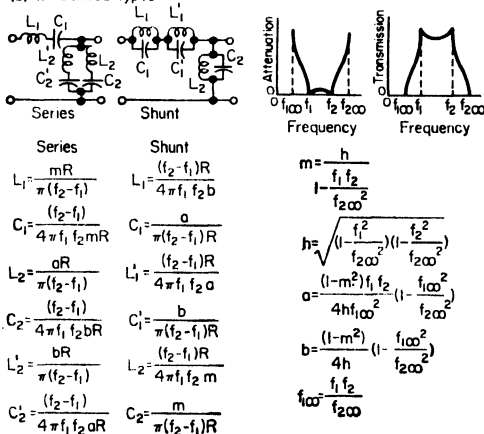
Series	Shunt
$L_1 = \frac{mR}{\pi(f_2 - f_1)}$	$L_1 = \frac{(f_2 - f_1)R}{4\pi f_1 f_2 b}$
$C_1 = \frac{(f_2 - f_1)}{4\pi f_1 f_2 mR}$	$C_1 = \frac{a}{\pi(f_2 - f_1)R}$
$L_2 = \frac{aR}{\pi(f_2 - f_1)}$	$L'_1 = \frac{(f_2 - f_1)R}{4\pi f_1 f_2 a}$
$C_2 = \frac{(f_2 - f_1)}{4\pi f_1 f_2 bR}$	$C'_1 = \frac{b}{\pi(f_2 - f_1)R}$
$L'_2 = \frac{bR}{\pi(f_2 - f_1)}$	$L_2 = \frac{(f_2 - f_1)R}{4\pi f_1 f_2 m}$
$C'_2 = \frac{(f_2 - f_1)}{4\pi f_1 f_2 aR}$	$C_2 = \frac{m}{\pi(f_2 - f_1)R}$

$$m = \frac{h}{1 - \frac{f_1 f_2}{f_{2\infty}^2}}$$

$$h = \sqrt{\left(1 - \frac{f_1^2}{f_{2\infty}^2}\right)\left(1 - \frac{f_2^2}{f_{2\infty}^2}\right)}$$

$$a = \frac{(1 - m^2) f_1 f_2}{4h f_{1\infty}^2}\left(1 - \frac{f_{1\infty}^2}{f_{2\infty}^2}\right)$$

$$b = \frac{(1 - m^2)}{4h}\left(1 - \frac{f_{1\infty}^2}{f_{2\infty}^2}\right)$$

$$f_{1\infty} = \frac{f_1 f_2}{f_{2\infty}}$$

Internal Junctions. The following terminations of the types of filter sections for which formulas are given in Sec. 66 may be joined together without impedance mismatches at the junction points:

Mid-series termination of constant-K type to mid-series termination of series-derived m type.

Mid-shunt termination of constant-K type to mid-shunt termination of shunt-derived m type.

Mid-series termination of constant-K, series-derived m type or shunt-derived m type to mid-series termination of another section of the same type.

Mid-shunt termination of constant-K, series-derived m type or shunt-derived m type to mid-shunt termination of another section of the same type.

NOTE: In the latter two cases, the values of m in the two sections to be joined, if they are of the m type, may be, and frequently are, different. Both sections must be of the same type and termination, however.

66. Filter-design Formulas. Formulas for calculating the capacitances and inductances of constant-K, series-derived m-type and shunt-derived m-type basic sections are given on pages 5-83 and 5-84. These are expressed in terms of R, the terminating resistances, the factor m, and the values of f_c, the cutoff frequency, and other critical frequencies. These factors must be predetermined on the basis of the filter requirements and the considerations outlined above.

Examples of Filter Design: 1. *Single-section Filter.* Required: High-pass single-section filter to be connected between resistance terminations of R = 1,000 ohms, with a cutoff frequency of 1,000 cycles and maximum attenuation occurring at 800 cycles.

To secure the attenuation peak at 800 cycles, an m-type filter section is required. Either the shunt- or series-derived type may be used. Choosing the latter, we have from the filter formulas II (b), Sec. 65,

$$
\begin{aligned}
f_1 &= 1{,}000 \text{ cycles} \\
f_{1\infty} &= 800 \text{ cycles} \\
R &= 1{,}000 \text{ ohms} \\
m &= 0.6 \\
C_1 &= 0.1325 \times 10^{-6} \text{ farad} \\
L_2 &= 0.1325 \text{ henry} \\
C_2 &= 0.298 \times 10^{-6} \text{ farad}
\end{aligned}
$$

From the considerations involving impedance matching at the end terminals, a mid-shunt termination facing each resistance termination is seen to be desirable for a series-derived section. Hence the structure of Fig. 89f is indicated. One full-series element (C_1) will be required, with a double-impedance shunt arm ($2L_2 + C_2/2$) at each end. The completed filter will then be as shown in Fig. 91.

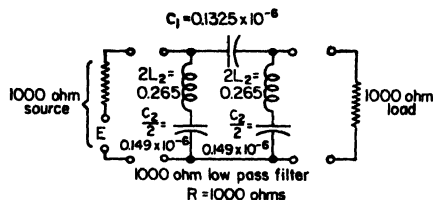



FIG. 91. Example of single-section filter.

2. *Multisection Composite Filter.* Required: Low-pass filter to be connected between resistance terminations of R = 600 ohms, with sharp cutoff at 1,000 cycles and high attenuation beyond.

There is no unique solution or "best" filter design for this problem. A large number of filters might be designed to meet these requirements, each of which would serve as well as any of the others. The relative merits of different designs will depend upon their economy of coils and capacitors in accomplishing the required results. One suitable design is shown here:

Let the input-end section be a half-section mid-series-derived m type, with its mid-shunt

termination facing the input to match impedances at that point. Let $m = 0.4$ for this half section to give a sharp cutoff.

This will be followed by a symmetrical full section of the series-derived m type, mid-series terminated, with $m = 0.75$; then a half section of the constant-K type with mid-series termination facing the full section and mid-shunt termination facing the end-terminating half section, which will be shunt-derived m type, with $m = 0.6$. The latter will have a

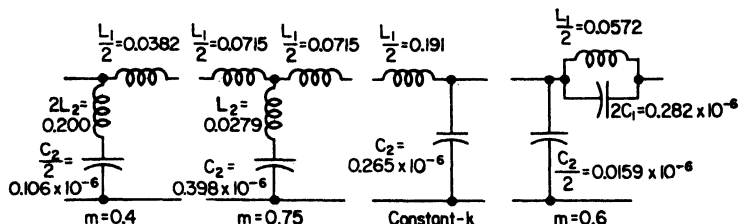



FIG. 92. Low-pass filter for use between 600 ohms with sharp cutoff at 1,000 cps.

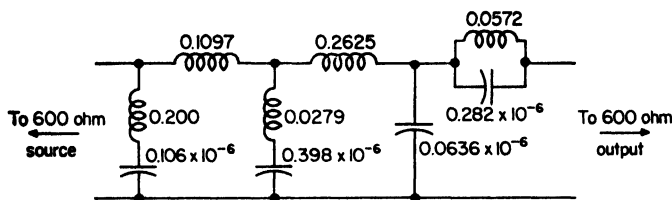



FIG. 93. Final filter as designed by Fig. 92.

mid-shunt termination facing the constant-K half section and a mid-series termination facing the output termination.

References

Transients:

Carter: "Simple Calculation of Electric Transients," Macmillan, 1945.
Frank, E.: "Pulsed Linear Networks," McGraw-Hill, 1945.
Kurtz, E. B., and G. F. Corcoran: "Introduction to Electric Transients," Wiley, 1935.
Pierce, G. W.: "Electric Oscillations and Electric Waves," McGraw-Hill, 1920.
Steinmetz: "Transient Electric Phenomena and Oscillations," McGraw-Hill, 1924.

A-c Circuits and Coupled Circuits:

Aiken: Two-mesh Coupled Tuned Filters, *Proc. IRE*, February, 1937.
Everitt, W. L.: "Communication Engineering," McGraw-Hill, 1937.
Pierce, G. W.: "Electric Oscillations and Electric Waves," McGraw-Hill, 1920.
"Radio Instruments and Measurements," *Nat. Bur. Standards Circ.* 74.
Terman, F. E.: "Radio Engineering," McGraw-Hill, 1947.

Transmission Lines and Networks:

Bode, H. W.: "Network Analysis and Feedback Amplifier Design," Van Nostrand, 1945.
Bronwell, A. B., and R. E. Beam: "Theory and Application of Microwaves," McGraw-Hill, 1947.
Brown: Characteristics of Unbalanced Overhead Transmission Lines, *RCA Broadcast News*, May, 1941.
Corrington, Murakami, and Sonnenfeldt: The Complete Specification of a Network by a Single Parameter, *RCA Rev.*, September, 1954.
Everitt, W. L., and G. E. Anner: "Communication Engineering," 3d ed., McGraw-Hill, 1956.
Foster, R. M.: A Reactance Theorem, *Bell System Tech. J.*, April, 1924.
Guillemin, E. A.: "Communication Networks," Wiley, 1935.
Johnson, K. S.: "Transmission Circuits for Telephonic Communication," Van Nostrand, 1925.
Karakash, J. J.: "Transmission Lines and Filter Networks," Macmillan, 1950.
Kennelly, A. E.: "Electric Lines and Nets," McGraw-Hill, 1928.
Mason, W. P., and R. A. Sykes: Use of Coaxial and Balanced Transmission Lines in Filters and Wide Band Transformers, *Bell System Tech. J.*, July, 1937.
Murakami and Corrington: Relation between Amplitude and Phase in Electrical Networks, *RCA Rev.*, December, 1948.
Paine: Graphical Solution of Transmission Lines, *Proc. IRE*, November, 1944.
Slater, J. C.: "Microwave Transmission," McGraw-Hill, 1942.

Sterba and Feldman: Transmission Lines for Short Wave Radio Systems, *Bell System Tech. J.*, July, 1932.
Weinberg, Louis: "Network Design by Use of Modern Synthesis Techniques and Tables," Technical Manual 427, April, 1956; "Additional Tables for Design of Optimum Ladder Networks," *Technical* Manual 434, Aug. 31, 1956, Hughes Aircraft Co.

Equalizers:

Brogle: Design of Reactance Equalizers, *Bell System Tech. J.*, October, 1949.
Zobel: Distortion Correction in Electrical Circuits with Constant Resistance Recurrent Networks, *Bell System Tech. J.*, July, 1928.

Wave Filters:

Campbell, G. A.: Physical Theory of Electric Wave Filters, *Bell System Tech. J.*, November, 1922.
Carson, J. R., and O. J. Zobel: Transient Oscillations in Electric Wave Filters, *Bell System Tech. J.*, July, 1923.
Giacolette: Optimum Resistive Terminations for Single Section Constant-*K* Ladder Type Filters, *RCA Rev.*, September, 1947.
Johnson, K. S., and J. E. Shea: Mutual Inductance in Wave Filters, with an Introduction on Filter Design, *Bell Sys em Tech. J.*, January, 1925.
Johnson, K. S.: "Transmission Circuits for Telephonic Communication," Van Nostrand, 1925.
Pierce, G. W.: "Electric Oscillations and Electric Waves," McGraw-Hill, 1920.
Shea, T. E.: "Transmission Networks and Wave Filters," Van Nostrand, 1925.
Zobel, O. J.: Theory and Design of Uniform and Composite Wave Filters, *Bell System Tech. J.*, January, 1923.
Zobel, O. J.: Transmission Characteristics of Electric Wave Filters, *Bell System Tech. J.*, October, 1924.
See also cumulative indexes of *Bell System Tech. J.*, *Proc. IRE*, and *RCA Rev.* for additional references.

CHAPTER 6

WAVE GUIDES AND CAVITY RESONATORS

By Harald Schutz[1]

1. Relationship between Wave Guides and Transmission Lines. A transmission line may be defined as a system of material boundaries forming a continuous path from one place to another and capable of directing energy along this path. The electromagnetic energy is carried along a transmission line in the form of guided electromagnetic waves. Transmission-line theory may be developed in terms of these traveling waves which propagate in either direction along the line.

Wave guides are included as transmission lines by the above definition. This is justified because many wave-guide phenomena are readily interpreted in terms of the

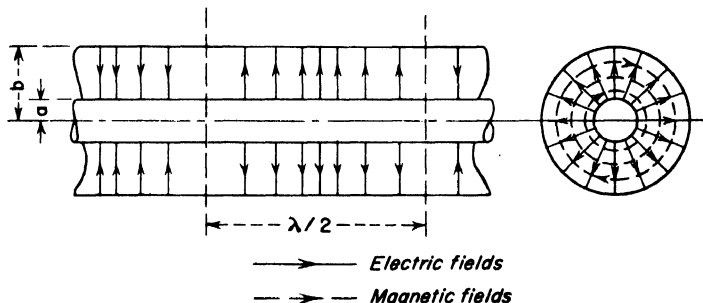



Fig. 1. Electric and magnetic fields for principal mode of propagation in coaxial line.

traveling-wave theory of transmission lines, and it is possible to draw equivalent transmission-line circuits for many wave-guide structures. Certain modifications to ordinary line theory are required for wave-guide transmission circuits. To determine these modifications, it is necessary to examine the nature of traveling waves on transmission lines.

A transmission line at ordinary radio frequencies consists of two or more separated conductors. When the line is energized, currents flow in these conductors and voltages exist between conductors. Electric and magnetic fields are found in the insulating medium separating the conductors. An example is the coaxial line shown in Fig. 1. The electric lines of force are shown as solid lines, and the magnetic lines of force as broken lines. The electric and magnetic lines of force are at right angles to each other, as is usually characteristic of electromagnetic waves. In the example, the electric and magnetic lines of force are also everywhere transverse to the direction of energy propagation, which is along the axis of the conductors. For this reason, the wave shown is called a *transverse electromagnetic wave*, abbreviated *TEM*.

TEM waves are used in all conventional transmission lines. If there are only two conductors, only one *TEM* wave is possible. If there are more than two conductors, more than one *TEM* wave may exist.[2]

[1] The Martin Co. Several sections of this chapter were written by Theodore Moreno and appeared in the Fourth Edition of this handbook, and the chapter as a whole leans heavily on his work for that edition.

[2] For example, in a two-wire overground transmission line, there is one wave in which the current

There may exist on a transmission line, in addition to the limited number of *TEM* waves, an infinite number of other types of waves, or higher modes of transmission. Each of these other modes has its own distinctive configuration of electric and magnetic fields. For a coaxial line, a few of the higher modes are shown in Fig. 2. The

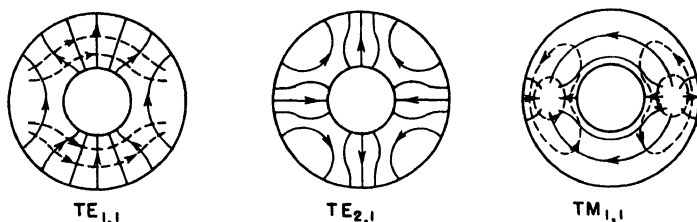



FIG. 2. Field configurations for some of the higher modes on coaxial line.

TEM wave, also called the *principal mode,* is distinguished from these higher modes in a number of ways. For example:

1. The transmission line acts as a high-pass filter for the higher modes, and the line will carry energy only if the frequency is above a cutoff frequency. This cutoff frequency is in general different for each higher mode. The higher modes may be excited at frequencies lower than their cutoff frequencies, but their strength will diminish very rapidly with increasing distance from the point of excitation.
2. The wavelength measured on the transmission line will be greater for the higher modes than for the principal mode.
3. The characteristic impedance and propagation constant of the line for the higher modes will differ from the values for the principal mode. The characteristic impedance is harder to define for the higher modes, as there no longer exists a unique value of voltage between conductors.

A transmission line requires two separated conductors to support a principal mode, but higher modes can carry energy on transmission lines that have only one conductor. For example, the coaxial line modes shown in Fig. 2 will reduce to the circular waveguide modes shown in Fig. 6 as the diameter of the center conductor is reduced to zero.

A wave guide is a transmission line not operating in a *TEM* mode. Most wave guides are incapable of supporting a *TEM* wave. They must therefore have less than two separated conductors. Normally a wave guide consists of a hollow conducting cylinder, of arbitrary cross section, with the electromagnetic fields confined to the inside of the cylinder. A dielectric rod may also act as a wave guide in the absence of any conductor.

Because of the phenomenon of cutoff that is peculiar to wave guides, the required dimensions of a wave guide for energy transmission are larger at lower frequencies. For most applications, wave guides are impractically large and bulky at frequencies below 1,000 to 2,000 Mc. On the other hand, conventional transmission lines, such as coaxial lines, must be restricted in size to keep the higher modes below cutoff[1] and become impractically small at frequencies much higher than 10,000 Mc. In the region between 1,000 and 10,000 Mc, both wave guides and coaxial lines are extensively used. Coaxial lines are smaller and lighter and, if cables are used, more flexible. On the other hand, wave guides offer lower attenuation, greater power-carrying capacity, and mechanical simplicity. A rectangular wave guide has a practical fre-

flows down one wire and returns in the other, and a second wave in which the current flows down both wires and returns through the ground.

[1] In a coaxial line, the cutoff wavelength of the first of the higher modes is given approximately by

$$\lambda_c \doteq \sqrt{\epsilon}\,\pi(b + a)$$

where b and a are the outer and inner radii of the line and ϵ is the dielectric constant of the medium separating the conductors.

quency range of less than two to one, while the band width of a coaxial line is much greater. A comparison of a typical wave guide, air dielectric coaxial line, and flexible cable is given in Table 1 for a frequency of 5,000 Mc.[1]

Table 1. Comparison of Transmission Lines for 5,000 Mc

Type of line	Rectangular wave guide	Rigid coaxial line	Flexible cable
Army-Navy type No.	RG-49/U	RG-76/U	RG-9B/U
Outside dimensions	2 × 1 in.	⅝ in. diam	0.425 in. diam
Conductor material	Brass	Brass	Inner conductor, 7/21 AWG silvered copper Outer conductor, double braid, inner silvered copper, outer copper
Surface finish	Silver	Silver	Silver
Dielectric	Air	Air	Polyethylene
Weight, lb per ft	1.40	0.292	0.158
Attenuation, db per ft	0.011	0.035	0.23
Recommended power rating	1.2 Mw	0.3 Mw	4,000 volts rms max, 66 watts continuous

2. Characteristics of Wave-guide Modes.[2] The higher modes which may carry energy in uniform wave guides with homogeneous dielectric are divisible into two classes: *TE* and *TM* modes, each infinite in number for all wave guides.

TE (Transverse Electric) Modes. The electric field is everywhere transverse to the direction of energy propagation, which is the direction of the guide axis. The magnetic field has transverse components, but in addition has components parallel to the axis. These modes are also called *H* waves

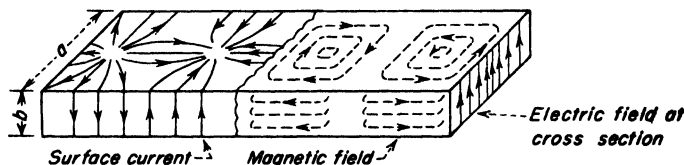



FIG. 3. Dominant mode, $TE_{1,0}$, in rectangular wave guide, showing electric and magnetic fields and conduction currents in walls.

TM (Transverse Magnetic) Modes. The magnetic field is everywhere transverse to the guide axis. The electric field has transverse components, but in addition has components parallel to the axis. These modes are also called *E* waves.

It is customary to identify individual modes by an additional subscript notation, the rules of which depend upon the guide configuration. For circular and rectangular cross sections, these rules are discussed in the following paragraphs.

For any wave guide, the mode of transmission that has the lowest cutoff frequency is called the *dominant mode.*

Rectangular Wave Guides.[3] Individual modes in rectangular wave guides are identified by giving the class of the transmission mode, followed by two numerical subscripts. For example, the dominant mode in rectangular wave guide, illustrated in Fig. 3, is the $TE_{1,0}$ mode. The first and second subscripts indicate the number of

[1] Moreno, T., Engineering Approach to Wave Guides, *Electronics*, **19** (5), 99, May, 1946.

[2] The four articles that form the basis of modern development of wave guides are: Southworth, G. C., Hyper-frequency Wave Guides—General Considerations and Experimental Results, *Bell System Tech. J.*, **15** (2), 284–309, April, 1936. Carson, J. R., S. P. Mead, and S. A. Schelkunoff, Hyper-frequency Wave Guides—Mathematical Theory, *Bell System Tech. J.*, **15** (2), 310–313, April, 1936. Barrow, W. L., Transmission of Electromagnetic Waves in Hollow Tubes of Metal, *Proc. IRE*, **24** (10), 1298, October, 1936. Chu, L. J., and W. L. Barrow, Electromagnetic Waves in Hollow Metal Tubes of Rectangular Cross Section, *Proc. IRE*, **26** (12), 1520, December, 1938.

[3] Chu and Barrow, *loc. cit.*

half-period variations in transverse field intensity along the larger and smaller cross-section dimensions, denoted by a and b, respectively. A few of the higher modes of transmission are illustrated in Fig. 4. The $TE_{1,0}$ mode in rectangular wave guide deserves special consideration, because it is by far the most widely used mode of transmission in practical wave guides. Also, a simple physical picture of the wave structure is possible, from which many of the basic formulas that apply to all types of wave guides may be developed.

The fields inside the wave guide may be considered to result from the addition of two plane waves, whose electric fields are parallel to the side walls of the guide.[1]

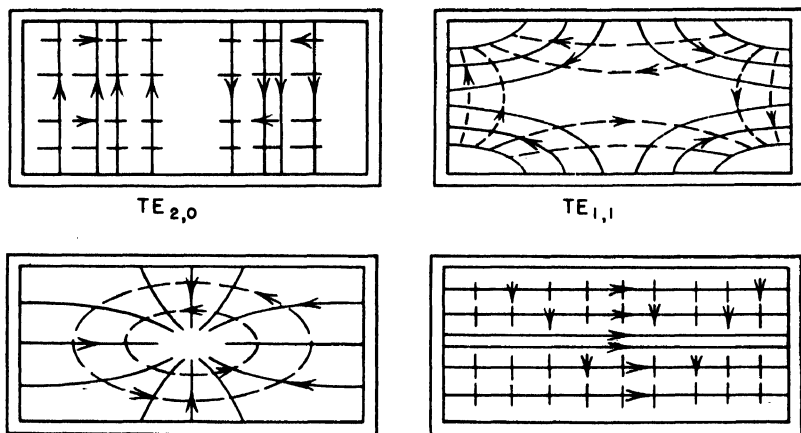



FIG. 4. Some of the higher modes of transmission in rectangular wave guide.

Each of these waves advances at an angle with respect to the axis of the guide, and each is the reflection of the other from the side walls of the guide. For the tangential electric field to be zero at the conducting side walls, the two plane waves must be out of phase at the walls; *i.e.*, the positive crest of one wave coincides with the negative crest of the other. This is illustrated in Fig. 5. Maximum electric field intensities are found at the center of the guide, where the two traveling waves are in phase and reinforce each other.

When the frequency is much higher than the cutoff frequency, the two component plane waves travel nearly parallel to the axis of the guide. As the frequency decreases toward cutoff, the angle between the guide axis and the direction of propagation of the component waves increases. At cutoff, this angle reaches 90 deg, and the two component waves are reflected back and forth between the side walls with no propagation of energy down the guide. At cutoff, the distance between the side walls is one-half a free-space wavelength. For the $TE_{1,0}$ mode in rectangular wave guide, the cutoff wavelength λ_c is related to the larger dimension of the wave guide a by

$$\lambda_c = 2a \tag{1}$$

The distance between positive and negative crests measured along the axis of the guide is seen by Fig. 5 to be greater than the distance between positive and negative crests of the component plane waves. This leads to the very important conclusion that the apparent wavelength in a wave guide, which is the distance between surfaces of equivalent phase, is greater than the free-space wavelength at the same frequency. If a wave guide is used as a Lecher wire system to measure wavelength, this correction must be made.

[1] Page, L., and N. I. Adams, Jr., Electromagnetic Waves in Conducting Tubes, *Phys. Rev.*, **52** (18), Sept. 15, 1937. Also see H. H. Skilling, Physical Behavior of Wave Guides, *Electronics*, **16** (3), 76, March, 1943, which gives a less mathematical but more graphical development.

More complex modes in rectangular wave guide may be constructed by the superposition of more than two plane waves. The method is applicable, in fact, to wave guides of any cross section, although in general an infinite number of plane waves are required and the calculation becomes tedious.[1]

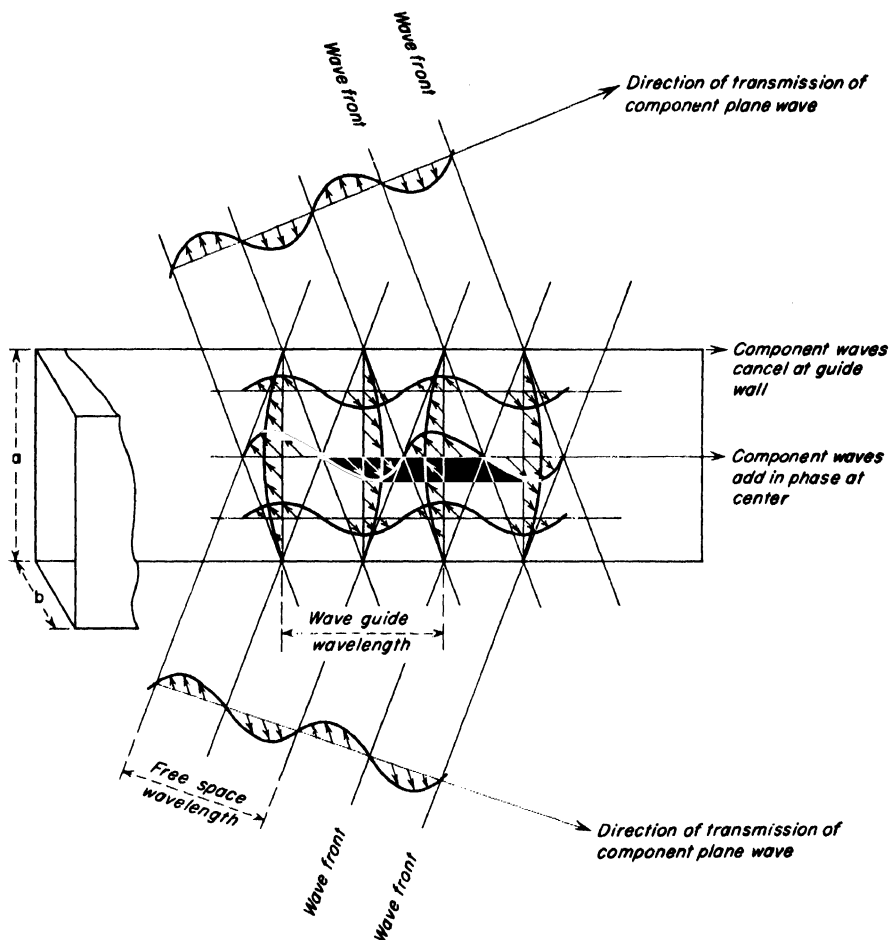



FIG. 5. Construction of the dominant mode in rectangular wave guide by the superposition of two plane waves traveling at an angle with respect to the guide walls.

3. Phase and Group Velocity. The component plane waves, which add to give the fields inside the guide, travel at the speed of light (3×10^{10} cm per sec). Since the wave-guide wavelength λ_g is greater than the wavelength of the component waves λ, this gives rise to a *phase velocity* in the wave guide v_p which exceeds the velocity of light c by the ratio v_p/c. The rate of energy propagation down the guide is not equal to the phase velocity, however, but is instead less than the velocity of light, because the component waves which carry the energy travel at an angle with respect to the guide axis. The rate of energy propagation along the axis is equal to the projection

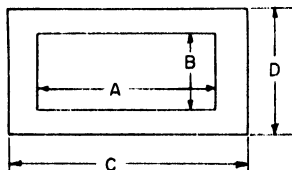
[1] Mackeown, S. S., and J. W. Miles, The Plane Wave Resolution of Guided Waves, *Proc. IRE*, **33** (11), 805, November, 1945.

upon the axis of the velocity vector of the component waves. This rate of energy propagation is known as the *group velocity* v_g. The group velocity and the phase velocity are related to the velocity of light by

$$v_p v_g = c^2 \tag{2}$$

This relationship holds for any mode in any air-filled wave guide.

Table 2. Dimensions, Tolerances, and Frequency Range for Rectangular Wave Guides

EIA designation	Frequency range (1,000 Mc), for dominant (TE_{10}) mode	Dimensions, in.								Max inner radius†
		Inner			Outer			Wall thickness		
		A	*B*	Tolerance	*C*	*D*	Tolerance	Nominal	Deviation from mean	
WR770	0.96–1.45	7.700	3.850	± .005	7.950	4.100	± .005	0.125	± .005	3/64
WR650*	1.12–1.70	6.500	3.250	± .005	6.660	3.410	± .005	0.080	± .005	3/64
WR510	1.45–2.20	5.100	2.550	± .005	5.260	2.710	± .005	0.080	± .005	3/64
WR430*	1.70–2.60	4.300	2.150	± .005	4.460	2.310	± .005	0.080	± .005	3/64
WR340*	2.20–3.30	3.400	1.700	± .005	3.560	1.860	± .005	0.080	± .005	3/64
WR284*	2.60–3.95	2.840	1.340	± .005	3.000	1.500	± .005	0.080	± .005	3/64
WR229	3.30–4.90	2.290	1.145	± .005	2.418	1.273	± .005	0.064	± .005	3/64
WR187*	3.95–5.85	1.872	0.872	± .005	2.000	1.000	± .005	0.064	± .005	1/32
WR159	4.90–7.05	1.590	0.795	± .004	1.718	0.923	± .004	0.064	± .005	1/32
WR137*	5.85–8.20	1.372	0.622	± .004	1.500	0.750	± .004	0.064	± .005	1/32
WR112*	7.05–10.00	1.122	0.497	± .004	1.250	0.625	± .004	0.064	± .005	1/32
WR90*	8.20–12.40	0.900	0.400	± .003	1.000	0.500	± .003	0.050	± .005	1/32
WR75	10.00–15.00	0.750	0.375	± .003	0.850	0.475	± .003	0.050	± .005	1/32
WR62*	12.40–18.00	0.622	0.311	± .0025	0.702	0.391	± .003	0.040	± .005	1/64
WR51	15.00–22.00	0.510	0.255	± .0025	0.590	0.335	± .003	0.040	± .005	1/64
WR42*	18.00–26.50	0.420	0.170	± .0020	0.500	0.250	± .003	0.040	± .005	1/64
WR34	22.00–33.00	0.340	0.170	± .0020	0.420	0.250	± .003	0.040	± .005	1/64
WR28*	26.50–40.00	0.280	0.140	± .0015	0.360	0.220	± .002	0.040	± .003	1/64
WR22*	33.00–50.00	0.224	0.112	± .0010	0.304	0.192	± .002	0.040	± .003	0.010
WR19	40.00–60.00	0.188	0.094	± .0010	0.268	0.174	± .002	0.040	± .003	0.010
WR15*	50.00–75.00	0.148	0.074	± .0010	0.228	0.154	± .002	0.040	± .003	0.008
WR12*	60.00–90.00	0.122	0.061	± .0005	0.202	0.141	± .002	0.040	± .003	0.006
WR10	75.00–110.00	0.100	0.050	± .0005	0.180	0.130	± .002	0.040	± .003	0.006

* Military standard.

† For all sizes: Min outer radius 1/64 in. Max outer radius 1/32 in.

The relationship between free-space wavelength λ, wave-guide wavelength λ_g, and cutoff wavelength λ_c for any mode in any air-filled wave guide is given by

$$\lambda_g = \frac{\lambda}{\sqrt{1 - (\lambda/\lambda_c)^2}} \tag{3}$$

For the important $TE_{1,0}$ mode in rectangular wave guide, the cutoff wavelength is

given by Eq. (1), and Eq. (3) reduces to

$$\lambda_g = \frac{\lambda}{\sqrt{1 - (\lambda/2a)^2}} \tag{4}$$

For all TE and TM modes in rectangular wave guide, the cutoff wavelength is given by

$$\lambda_c = \frac{2}{\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}} \tag{5}$$

In Eq. (5), m and n are the subscripts denoting the particular mode under consideration (*e.g.*, $TE_{m,n}$), and also indicate the number of half-period variations in transverse field intensity along the a and b cross-section dimensions of the guide, respectively.

If a rectangular guide is to be above cutoff for the dominant $TE_{1,0}$ mode, but below cutoff for all other modes,

$$a > \frac{\lambda}{2} < \lambda$$

$$b < \frac{\lambda}{2}$$

Various standard sizes of rectangular wave guide for broad-band use have been recommended by the EIA Subcommittee on Microwave Transmission Lines. These are given in Table 2, taken from *Electronics*, p. 111, June, 1949, along with the frequency range for which each of the guides is recommended.

4. Circular Wave Guides.[1] A number of possible modes of transmission in circular wave guides are shown in Fig. 6. For any mode of transmission in a circular wave guide, the transverse fields may be resolved into two components: tangential and radial. Both of these components vary periodically along a circular path concentric with the wall, and both vary in a manner related to a Bessel function of the first kind of order m along a radius. Any particular mode is identified by the notation $TE_{m,n}$ or $TM_{m,n}$, where m is the total number of full-period variations of either component of field along a circular path concentric with the wall, and n is one more than the total number of reversals of sign of either component of field along a radial path.

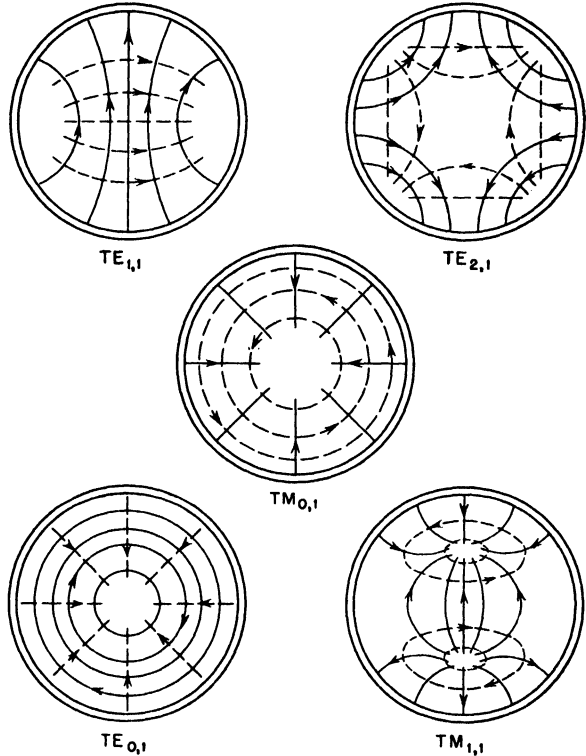



Fig. 6. Some transmission modes in circular wave guides.

[1] Southworth, *loc. cit.* Carson, Mead, and Schelkunoff, *loc. cit.* Barrow, *loc. cit.*

The dominant mode in circular wave guide, that with the longest cutoff wavelength, is the $TE_{1,1}$ mode, which corresponds to the $TE_{1,0}$ mode in rectangular guide. This mode is shown in Fig. 6. Another mode that is of considerable interest is the $TM_{0,1}$, or circular magnetic mode. This is the lowest mode with circular symmetry. A mode of considerable theoretical interest is the $TE_{0,1}$ mode, which has anomalous attenuation characteristics discussed below.

The cutoff wavelength for all modes in circular wave guide depends upon the ratio of diameter to wavelength. For the $TE_{m,n}$ modes, the cutoff wavelength is given by

$$\lambda_c = \frac{2\pi a}{u_{m,n}'} \tag{6}$$

where a is the inner radius of the guide. The constant $u_{m,n}'$ is the nth root of the equation $J_m'(u) = 0$, where the prime denotes differentiation. Some of the lower values of $u_{m,n}'$ are

$u_{01}' = 3.832$	$u_{02}' = 7.016$	
$u_{11}' = 1.841$	$u_{12}' = 5.332$	$u_{13}' = 8.536$
$u_{21}' = 3.054$	$u_{22}' = 6.706$	
$u_{31}' = 4.201$	$u_{32}' = 8.016$	

For the $TM_{m,n}$ modes, the cutoff wavelength is given by

$$\lambda_c = \frac{2\pi a}{u_{m,n}} \tag{7}$$

where a is the inner radius of the guide and $u_{m,n}$ is the nth root of the equation $J_m(u) = 0$. Some of the lower values of $u_{m,n}$ are

$u_{01} = 2.405$	$u_{02} = 5.520$	$u_{03} = 8.654$
$u_{11} = 3.832$	$u_{12} = 7.016$	
$u_{21} = 5.136$	$u_{22} = 8.417$	
$u_{31} = 6.380$		

The cutoff frequency of the $TM_{0,1}$ mode is only 30 per cent higher than the cutoff frequency of the $TE_{1,1}$ mode, and the frequency range over which only the dominant $TE_{1,1}$ mode is above cutoff is correspondingly limited. To operate a wave guide over a wider frequency band in any mode of transmission, the means of excitation must be chosen so as to excite only the desired mode, or else mode dampers must be employed. These mode dampers may consist of grids of wire designed to short-circuit the undesired modes while leaving the desired mode unaffected. Methods of selective mode excitation are discussed later.

Fig. 7. Modes in elliptical guide and corresponding modes in circular wave guide.

5. Elliptical Wave Guides. The inevitable deformations encountered in circular wave guides result in an equivalent ellipticity, and the properties of elliptical

wave guides are, therefore, of practical interest.[1] A few of the possible modes in elliptical wave guide are shown in Fig. 7, along with the corresponding modes in circular wave guide. In general, if a round wave guide is deformed, each mode will split into two modes of orthogonal polarization with different phase velocities and different attenuations. Only those modes with circular symmetry (*e.g.*, $TE_{0,1}$ and $TM_{0,1}$) do not have this instability.

Other cross sections for wave guides have been the subject of some theoretical investigation but have found limited practical application to date. One of the more interesting designs is the H-shaped cross section[2] shown in Fig. 8. A wave guide of this design has over-all dimensions that are relatively small compared to the cutoff wavelength. Also, the frequency separation between cutoff for the dominant mode of transmission and that for the higher mode is comparatively large.

6. Characteristic Impedance of Wave Guides. The characteristic impedance of ordinary two-wire transmission lines may be defined as the input impedance to a line of infinite length and may be expressed as the ratio of voltage E between conductors to the current I flowing in each conductor ($Z_0 = E/I$) for a transmission line free from standing waves. It is not possible to extend this definition to wave guides, because there are no separated conductors between which a unique value of voltage exists. In an ordinary transmission line, it is possible also to define characteristic impedance in terms of the net power flow W into a matched line ($Z_0 = W/I^2$ or $Z_0 = E^2/W$). The net power flow into an infinite wave guide may be calculated when the field strengths at a cross section are given. If a certain logical value of voltage or current is then chosen as reference, it is possible to define a characteristic impedance of the guide in terms of the reference voltage or current and the power input. The actual numerical value of the impedance will depend upon precisely what voltage or current was chosen for reference. The ratio between values of characteristic impedance obtained by different choices of reference voltage or current will be numerically constant, however, and independent of changing frequency or guide dimensions.

FIG. 8. H-shaped wave guide.

For example, with the $TE_{1,0}$ mode in rectangular wave guide, if the maximum voltage between top and bottom guide faces is chosen as the reference voltage, the characteristic impedance of the guide becomes[3]

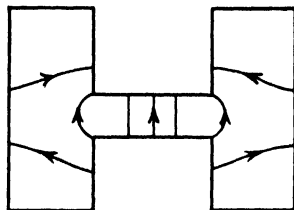
$$Z_0 = 377 \frac{\lambda_g}{\lambda} \frac{2b}{a} \tag{8}$$

If the total axial current in the top or bottom face is chosen as the reference current, the characteristic impedance will differ from the above expression only by the numerical factor of $\pi^2/16$.

For many applications, the numerical value of the characteristic impedance is of little importance, as the impedances of the associated elements in the circuit may be normalized with respect to the guide impedance. It is, however, important to adjust impedances of various transmission-line and wave-guide components to match them for most efficient power transmission and minimum reflection at the junctions. Transformers must be designed and adjusted. They take many physical shapes. A number of impedance and admittance charts have been devised, and some of them have found widespread application.

[1] Schelkunoff, S. A., Note on Certain Guided Waves in Slightly Noncircular Tubes, *J. Applied Phys.*, **9** (7), 484, July, 1938. Chu, L. J., Electromagnetic Waves in Elliptic Hollow Pipes of Metal, *J. Applied Phys.*, **9** (9), 583, September, 1938.

[2] Ramo, S., and J. R. Whinnery, "Fields and Waves in Modern Radio," p. 349, Wiley, 1944.

[3] Schelkunoff, S. A., Impedance Concept in Wave Guides, *Quart. Applied Math.*, **2** (1), 1–15, April 1944.

The relative admittance at any point z of a transmission line can be related to that at point z_0 by

$$Y'(z) = \frac{j + Y'(z_0) \cot k(z_0 - z)}{\cot k(z_0 - z) + jY'(z_0)} \tag{8a}$$

where $k = 2\pi/\lambda_g$.

Equation (8a) can be rewritten as a fundamental impedance relation

$$Z'(z) = \frac{j + Z'(z_0) \cot k(z_0 - z)}{\cot k(z_0 - z) + jZ'(z_0)} \tag{8b}$$

The similarity between Eqs. (8a) and (8b) indicates the validity of a duality principle for transmission-line equations.

P. H. Smith of the Bell Telephone Laboratories developed the most frequently used system of transmission-line charts for microwave use.[1] In recent years graphical methods of presenting impedance and admittance relations have been developed further. Examples are the hyperbolic protractor of Deschamps[2] and the work of Wheeler.[3]

7. Power-carrying Capacity of Wave Guides. The power-carrying capacity depends upon the maximum electric field strength that can exist without breakdown. Experimental evidence indicates that at microwave frequencies, this maximum field strength is 30,000 volts per cm under standard sea-level conditions of temperature, pressure, and humidity.

For rectangular wave guide operating in the $TE_{1,0}$ mode, the maximum power P is related to the maximum field strength E_{max} by

$$\frac{P}{E_{max}^2} = 6.63 \times 10^{-4} ab \left(\frac{\lambda}{\lambda_g}\right) \tag{9}$$

If E_{max} is given in volts per centimeter, the inside dimensions of the guide, a and b, should be given in centimeters for the power to be in watts. The free-space wavelength is λ, the guide wavelength λ_g. The maximum field intensity occurs parallel to the narrower dimension of the guide, midway between the side walls, and is independent of the distance from the wide faces of the guide.

For circular wave guides operating in the dominant $TE_{1,1}$ mode the relation between maximum power and maximum field strength is

$$\frac{P}{E_{max}^2} = 1.99 \times 10^{-3} a^2 \left(\frac{\lambda}{\lambda_g}\right) \tag{10}$$

where a = radius of guide, cm
λ = free-space wavelength
λ_g = wave-guide wavelength

Maximum field strength is at the center of the guide.

For the $TM_{0,1}$ mode in circular wave guide, there are two cases:

Case 1—$a/\lambda < 0.761$:

$$\frac{P}{E_{max}^2} = 7.69 \times 10^{-3} \frac{a^4}{\lambda^2} \left(\frac{\lambda}{\lambda_g}\right) \tag{11}$$

The maximum field intensity is at the center of the guide.

[1] Smith, P. H., Transmission Line Calculator, *Electronics*, January, 1939, p. 29; Improved Transmission Line Calculator, *Electronics*, January, 1944, p. 130.

[2] Deschamps, G., A New Chart for the Solution of Transmission Line and Polarization Problems, *Trans. IRE*, **MTT-1,** March, 1953.

[3] Wheeler, H. A., Geometric Relations in a Circle Diagram of Transmission-line Impedance, Wheeler Monographs, 4, July, 1948.

Case 2—$a/\lambda > 0.761$:

$$\frac{P}{E_{\max}^2} = 3.33 \times 10^{-3} a^2 \left(\frac{\lambda_g}{\lambda}\right) \tag{12}$$

The maximum field intensity is at a radius of $0.765a$ and is independent of angle. All dimensions in the above formulas should be given in centimeters for the power to be in watts.

The power-carrying capacity of a wave guide will diminish with altitude approximately as shown in Fig. 9.

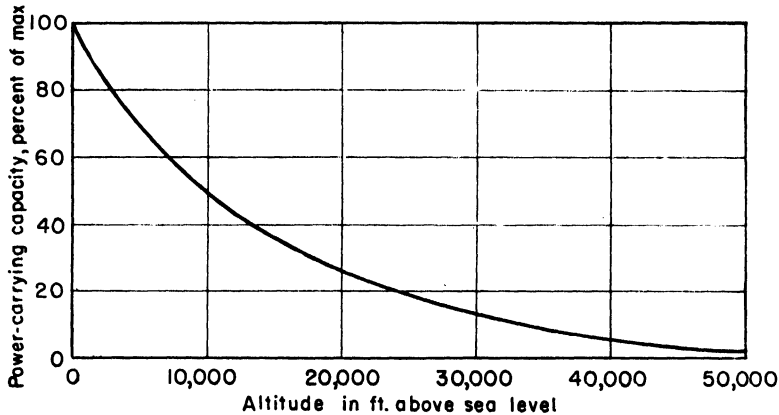
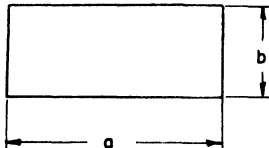



Fig. 9. Decrease in power-handling capacity at increasing altitude.

8. Attenuation in Wave Guides. At frequencies below cutoff, the attenuation in any wave guide is very rapid, even for perfectly conducting walls, as the guide is unable to support traveling waves. At frequencies higher than cutoff, the guide can support traveling waves, but they are attenuated because of losses in the conducting walls and in the dielectric that fills the guide. For air-filled guides, the dielectric losses are usually negligible, but if dielectrics other than air are used, the dielectric losses are often greater than the conductor losses.

9. Conductor Losses. The conductor losses in a wave guide are governed in part by skin-depth considerations, as the current that flows in the guide walls is concentrated near the inner surface. The attenuation will vary as the square root of the resistivity of the material which forms the conducting walls.

As the frequency increases beyond the cutoff frequency, the attenuation drops rapidly from a very high value at cutoff to a broad minimum, and then increases slowly with increasing frequency. This is the behavior for all modes in all wave guides except the family of circular electric modes (*e.g.*, $TE_{0,1}$) in circular wave guide. These modes have anomalous characteristics that are discussed in a succeeding paragraph.

The formulas that are given below for wave-guide attenuation indicate that the attenuation approaches infinity as the frequency approaches cutoff. However, the approximations under which these equations were developed are not valid in the region immediately adjacent to cutoff, for there is actually a smooth transition between the approximate results that are given below for the above-cutoff and below-cutoff regions.

10. Attenuation in Rectangular Wave Guides. Formulas for the attenuation in rectangular, copper, air-filled wave guide are given in Table 3. The inner dimensions of the guide, a and b, are in inches. The larger dimension is a. If some metal other than copper is used as a conductor, the attenuation given by Table 3 should be multiplied by the square root of the resistivity of the metal compared to that of copper.

Table 3. Attenuation in Rectangular, Copper, Air-filled Wave Guides

f = frequency
f_c = cutoff frequency
a and b in inches

Mode	α, Db/Ft
$TE_{1,1}$	$\frac{0.01107}{a^{3/2}}\left[\frac{\frac{1}{2}\frac{a}{b}\left(\frac{f}{f_c}\right)^{3/2}+\left(\frac{f}{f_c}\right)^{-1/2}}{\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right]$
$TE_{2,0}$	$\frac{0.01565}{a^{3/2}}\left[\frac{\frac{1}{2}\frac{a}{b}\left(\frac{f}{f_c}\right)^{3/2}+\left(\frac{f}{f_c}\right)^{-1/2}}{\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right]$
$TE_{1,1}$	$\frac{0.01107}{a^{3/2}}\left\{\frac{\frac{a}{b}\left(1+\frac{a}{b}\right)\left(\frac{f}{f_c}\right)^{3/2}+\left[1+\left(\frac{a}{b}\right)^2\right]\left(\frac{f}{f_c}\right)^{-1/2}}{\left[1+\left(\frac{a}{b}\right)^2\right]^{3/4}\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right\}$
$TM_{1,1}$	$\frac{0.01107}{a^{3/2}}\left\{\frac{\left[1+\left(\frac{b}{a}\right)^3\right]\left(\frac{f}{f_c}\right)^{3/2}}{\left[1+\left(\frac{b}{a}\right)^2\right]^{3/4}\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right\}$

The variation of attenuation with frequency in a typical rectangular wave guide is given in Fig. 10 for some of the lower modes of propagation. In Fig. 11, the variation of attenuation with frequency is given for the dominant $TE_{1,0}$ mode for a number of widely used guide sizes over the frequency range for which the guide is recommended.

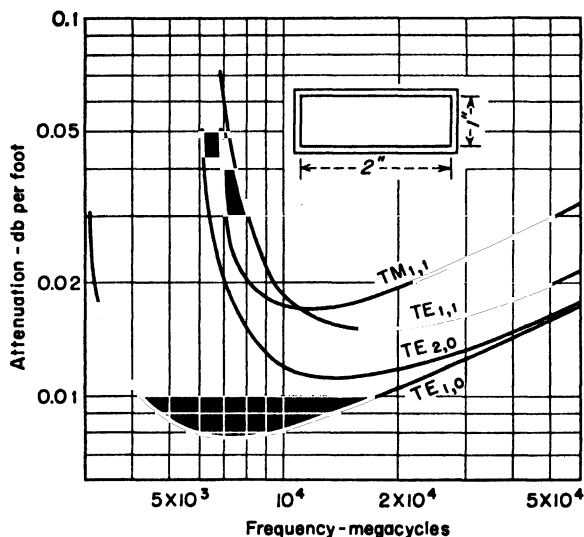
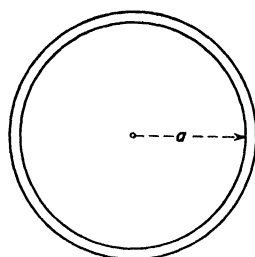



Fig. 10. Attenuation in typical rectangular, air-filled copper wave guide.

11. Circular Wave Guides. Attenuation formulas for various modes in circular, copper wave guide are given in Table 4. The guide radius a should be given in inches. The variation of attenuation with frequency for each of the above modes in a copper wave guide of 2 in. inside diameter is given in Fig. 12. Note that the attenuation of the $TE_{0,1}$ mode decreases without limit with increasing frequency.[1] Experimental verification of this is lacking, and this anomalous attenuation charac-

[1] Kemp, J., Anomalous Attenuation in Waveguides, *Wireless Engr.*, **23** (275), 211, August, 1946.

Table 4. Attenuation in Circular, Copper Wave Guides

f = frequency
f_c = cutoff frequency
Radius a in inches

Mode	α, Db/Ft
$TE_{1,1}$ (dominant)	$\frac{0.00423}{a^{3/2}}\left[\frac{\left(\frac{f}{f_c}\right)^{-1/2}+\frac{1}{2.38}\left(\frac{f}{f_c}\right)^{3/2}}{\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right]$
$TM_{0,1}$ (circular magnetic)	$\frac{0.00485}{a^{3/2}}\left[\frac{\left(\frac{f}{f_c}\right)^{3/2}}{\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right]$
$TE_{0,1}$ (circular electric)	$\frac{0.00611}{a^{3/2}}\left[\frac{\left(\frac{f}{f_c}\right)^{-1/2}}{\sqrt{\left(\frac{f}{f_c}\right)^2-1}}\right]$

teristic is lost if the guide is elliptical. Nevertheless, this mode offers the lowest attenuation of any of the lower wave-guide modes, provided troublesome effects from other modes which are above cutoff can be eliminated. The changes in attenuation caused by small amounts of ellipticity are small in magnitude unless the frequency is far above cutoff.

12. Comparison of Theoretical and Measured Attenuations. The values of attenuation given in the preceding paragraphs are all derived from theoretical considerations. Because the current in a wave guide is concentrated at the inner surface of the guide, the condition of this surface affects the attenuation, which is usually higher than predicted theoretically. The difference will generally be small at the low end of the microwave spectrum, but at higher frequencies surface roughness will increase the attenuation by increasing the path length which the surface current must travel. For example, at 25,000 Mc, even a well-machined surface will frequently have only two-thirds the theoretical conductivity of an ideal surface. Drawn brass tubing usually approaches rather closely the theoretical values at most microwave frequencies.

FIG. 11. Theoretical attenuation of copper wave guide over the recommended operating frequency range for the dominant mode. Dimensions are outside measurements and wall thickness.

The skin depth in most wave guides is so small that if the surface of the guide is plated to a thickness of a thousandth of an inch or less, essentially all of the surface

current will flow in the plated surface metal. But because of the porosity of the plated surfaces, the attenuation will be somewhat higher than expected from the l-f resistivity of the plated metal. The condition of the surface before plating and the plating technique will have considerable effect upon the attenuation. For these reasons, attenuation in plated wave guides varies considerably from sample to sample, and at 10,000 Mc will frequently be half again as great as the theoretical value.

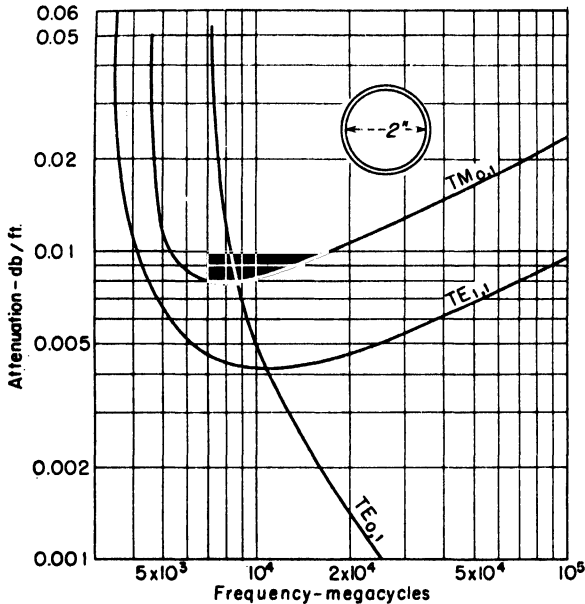



Fig. 12. Attenuation in typical circular, air-filled copper wave guide.

If the inside of a wave guide is coated with a thin protective coating of high-resistance material, such as protective lacquer, the current will flow under this layer in the base metal, and the attenuation will not be greatly affected.

13. Attenuation in Wave Guides below Cutoff. At frequencies below cutoff, wave guides are unable to support traveling waves which carry energy down the guide, and if modes are excited by some launching mechanism, the field intensity will diminish very rapidly with increasing distance from the point of excitation.[1] The input impedance of a wave guide below cutoff is a pure reactance.

For any one mode in the wave guide, the field strength as a function of the distance from the point of excitation will vary as

$$E = E_0\epsilon^{-\alpha z} \tag{13}$$

where E_0 is the initial amplitude at the point of excitation. The attenuation is given by

$$\alpha = 8.69\sqrt{\left(\frac{2\pi}{\lambda_c}\right)^2 - \epsilon\left(\frac{2\pi}{\lambda}\right)^2} \quad \text{db/unit length} \tag{14}$$

where λ_c = cutoff wavelength of air-filled guide
λ = wavelength in free space at frequency of excitation
ϵ = dielectric constant of dielectric medium inside wave guide

At frequencies much lower than the cutoff frequency, the attenuation is very nearly independent of frequency, approaching a limiting value at low frequencies of

$$\alpha = \frac{54.6}{\lambda_c} \quad \text{db/unit length} \tag{15}$$

[1] Linder, E. G., Attenuation of Electromagnetic Fields in Pipes Smaller than the Critical Size, *Proc. IRE*, **30** (9), 412, September, 1942.

For the two modes of greatest interest in circular wave guide, the modes that are commonly used in wave-guide below-cutoff attenuators, the attenuations are as follows:

For the $TE_{1,1}$ mode,

$$\alpha = 8.69 \sqrt{\left(\frac{1.841}{a}\right)^2 - \left(\frac{2\pi}{\lambda}\right)^2} \quad \text{db/unit length} \tag{16}$$

For the $TM_{0,1}$ mode,

$$\alpha = 8.69 \sqrt{\left(\frac{2.405}{a}\right)^2 - \left(\frac{2\pi}{\lambda}\right)^2} \quad \text{db/unit length} \tag{17}$$

where a is the guide radius and λ the free-space wavelength.

14. Wave Guides Filled with Dielectric Material. If a wave guide is partially or completely filled with some dielectric material, the characteristics of the guide will be modified. In dealing with these modifications, it is convenient to regard the dielectric constant ϵ of the material as complex and of the form

$$\epsilon = \epsilon' - j\epsilon'' \tag{18}$$

The loss tangent of the dielectric is then defined by

$$\tan \delta = \frac{\epsilon''}{\epsilon'} \tag{19}$$

For all but high-loss dielectrics, the loss tangent is very nearly equal to the power factor, and the two terms are frequently used interchangeably.

For all modes in wave guides, the cutoff wavelength of a dielectric-filled guide λ_c' is related to the cutoff wavelength of the same guide when air-filled λ_c by

$$\frac{\lambda_c'}{\lambda_c} = \sqrt{\epsilon} \tag{20}$$

The wave-guide wavelength in dielectric-filled guide for all modes is given by

$$\lambda_g = \frac{\lambda}{\sqrt{\epsilon' - (\lambda/\lambda_c)^2}} \tag{21}$$

where λ_c is the cutoff wavelength in the same guide when air-filled.

The standing-wave ratio associated with the reflection from an air-dielectric interface normal to the axis of the wave guide is, for TE modes,

$$\text{SWR} = \sqrt{\frac{1 - (\lambda/\lambda_c)^2}{\epsilon' - (\lambda/\lambda_c)^2}} \tag{22}$$

and, for TM modes,

$$\text{SWR} = \frac{1}{\epsilon'} \sqrt{\frac{\epsilon' - (\lambda/\lambda_c)^2}{1 - (\lambda/\lambda_c)^2}} \tag{23}$$

This reflection can be minimized by tapering the interface in the axial direction.

The attenuation in a below-cutoff wave guide filled with dielectric is given in Eq. (14).

In a wave guide above cutoff filled with dielectric, the attenuation resulting from dielectric losses is given by

$$\alpha = 27.2 \frac{\epsilon''}{\lambda} \frac{1}{\sqrt{\epsilon' - (\lambda/\lambda_c)^2}} \quad \text{db/unit length} \tag{24}$$

The conductor losses are also affected by the dielectric. To find the conductor losses in a dielectric-filled wave guide, it is necessary to modify the expressions in Tables 3 and 4 as follows: (1) In all expressions, replace f_c, the cutoff frequency of the air-filled guide, by $f_c/\sqrt{\epsilon}$, the cutoff frequency of the dielectric-filled guide; (2) multiply the resulting expression by $\epsilon^{1/4}$.

The total attenuation in a dielectric-filled wave guide is the sum of the attenuation resulting from dielectric losses and that resulting from conductor losses.

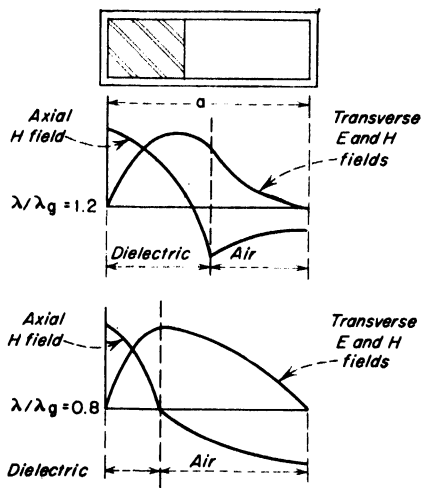



Fig. 13. Field intensities in a rectangular wave guide partially filled with a dielectric operating in the $TE_{1,0}$ mode.

15. Wave Guides Partially Filled with Dielectric.[1] If a wave guide is partially filled with a dielectric, its properties will, as expected, be intermediate between air-filled and dielectric-filled guides. The fields inside the guide tend to be pulled inside the medium with the higher dielectric constant. The distribution of the fields for a wave guide partially filled with dielectric is shown in Fig. 13, and it is apparent how the point of maximum electric field strength, normally in the center of the guide, has been pulled toward the dielectric.

Placing the dielectric inside the guide will increase the cutoff wavelength and reduce the wave-guide wavelength, but the magnitude of the effect will depend upon the location of the dielectric. The effects will be small, for example, if the dielectric is placed in a region of a weak electric field, as shown in Fig. 14, but will be much greater if the dielectric is placed in the center of the guide, in a region of strong field.

When the dielectric is tapered to minimize the reflection from a transition between dielectric-filled and air-filled guides, it is advisable to start and end the taper in the regions of the weak field. For example, with the $TE_{1,0}$ mode in rectangular wave guides, the best results are found with a taper constructed as shown in Fig. 15.

Dielectric in a region of weak electric field, small effect

Dielectric in a region of strong electric field, considerable effect

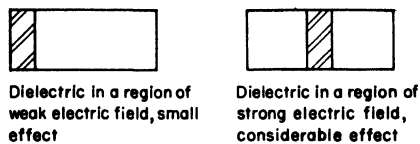



Fig. 14. A dielectric placed in a region of strong electric field will have a greater effect upon the properties of the wave guide than if it is placed in a region of weak field.

Fig. 15. Impedance discontinuity at the junction of air-filled wave guide and dielectric-filled wave guide can be minimized by tapering the dielectric.

16. Dielectric Wave Guides. It is possible for a dielectric cylinder immersed in a medium of lower dielectric constant to act as a wave guide in the absence of any conducting surface.[2] Such a wave guide without metal walls has many properties

[1] Pincherle, L., Electromagnetic Waves in Metal Tubes Filled Longitudinally with Two Dielectrics, *Phys. Rev.*, **66** (5, 6), Sept. 1 and 15, 1944.

[2] For example, see Ramo and Whinnery, *loc. cit.*

similar to more conventional wave guides; energy is carried in *TE* or *TM* modes, and the guide acts as a high-pass filter, with the cutoff frequency different for the different modes. But the electromagnetic fields are not confined to the inside of the dielectric cylinder, and a considerable fraction of the total energy carried by the guide is found in the fields that are external to the cylinder. This introduces a number of practical difficulties into the use of these dielectric wave guides. Any mechanical support for the guide interferes with the fields which are external to the dielectric cylinder, and constitutes a discontinuity in the transmission line. Furthermore, this discontinuity results not only in a partial reflection of the transmitted wave, but also in a partial loss of energy by radiation. A similar loss of energy by radiation will occur at a bend in the dielectric transmission line.

17. Obstacles and Discontinuities in Wave Guides. A metallic obstacle placed in a wave guide will in general affect the transmission of an electromagnetic wave through the guide. The obstacle will cause a partial reflection of the incident wave and, in addition, will excite other of the wave-guide transmission modes. If these are also above cutoff, part of the energy in the incident wave will be carried away from the discontinuity by these other modes of transmission. If the modes excited by the discontinuity are below cutoff, they will diminish rapidly in strength with increasing distance from the discontinuity and will draw no real power. They do contribute a reactive effect at the discontinuity, however.

Fig. 16. Equivalent susceptance of a thin symmetrical inductive window in a rectangular guide. Theoretical curves, experimental points.

The effect of an obstacle in a wave guide is equivalent to that of a discontinuity in a conventional transmission line, and it is convenient for engineering work to set up an

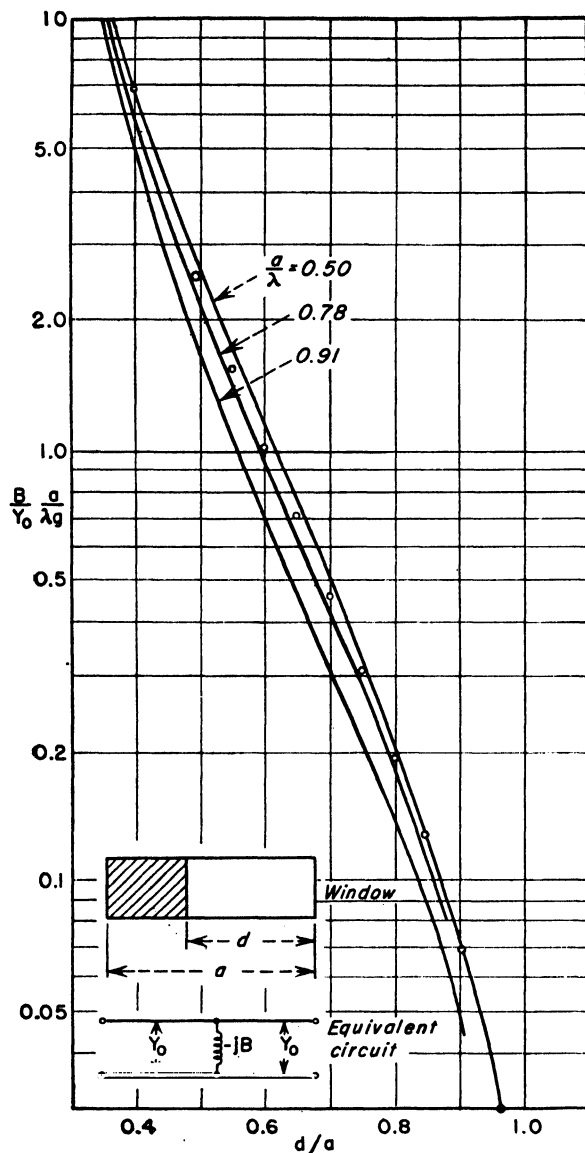



FIG. 17. Equivalent susceptance of thin asymmetrical inductive window in rectangular wave guide.

equivalent transmission line and lumped constant circuit for the wave-guide structure. This equivalent circuit will in general be a function of frequency.

If the obstacle has axial dimensions that are very small, *e.g.*, a thin metallic window that partially blocks the guide, the equivalent circuit will be a simple susceptance shunting the guide that is located precisely at the point of discontinuity.[1] If the obstacle is of appreciable thickness, a more complex equivalent circuit in the form of a tee or pi network is usually required, the precise circuit depending upon what planes in the wave guide are chosen for reference.

If two obstacles are placed close enough together for interaction of the below-cutoff higher-order modes, the equivalent circuits will be correspondingly modified. Interaction effects are usually small if the spacing is greater than a quarter-wave-guide wavelength.

The equivalent circuits given in this chapter are in general applicable when the wave guide is above cutoff for the dominant mode only.

18. Obstacles in Rectangular Wave Guide. *Thin Metallic Windows.*[2] A thin metallic window that partially blocks the guide has an equivalent circuit consisting of a susceptance shunting the guide at the location of the window. If the window is of the type shown in Fig. 16, the

[1] Whinnery, J. R., and H. W. Jamieson, Equivalent Circuits for Discontinuities in Transmission Lines, *Proc. IRE*, **32** (2), 98, February, 1944. Also Schelkunoff, S. A., "Electromagnetic Waves," p. 490, Van Nostrand, 1943.

[2] Miles, J. W., The Equivalent Circuit for a Plane Discontinuity in a Cylindrical Wave Guide, *Proc. IRE*, **34** (10), 728, October, 1946.

equivalent circuit is an inductive susceptance, whose magnitude is given in Fig. 16 as a function of the window dimensions. The curves are for infinitely thin windows, the experimental points for windows of 1/32 in. thickness at a wavelength of 6.0 cm. The effect of finite thickness is to increase the susceptance of the window, and series reactance terms in the equivalent circuit are also necessary if the thickness is considerable.

If the opening in the diaphragm is of the type shown in Fig. 17, the equivalent circuit is again a shunt inductive susceptance. The susceptance as a function of the dimensions is given in Fig. 17, along with some experimental results, taken with a window 1/32 in. thick at a wavelength of 6.0 cm.

If the window is of the type shown in Fig. 18, the equivalent circuit is a capacitive susceptance shunted across the guide. The theoretical susceptance for an infinitely thin window is given in Fig. 18. With a window of this type, it is very important for accurate results to take into account the finite thickness of the window. If the window thickness is finite but much less than a wavelength, its susceptance will be given by

FIG. 18. Equivalent susceptance of a thin capacitive window in rectangular wave guide. Theoretical curves for infinitely thin window.

$$\frac{B}{Y_0} = \frac{B_0}{Y_0} + \frac{2\pi t}{\lambda_g}\left(\frac{b}{d} - \frac{d}{b}\right) \tag{25}$$

where B_0/Y_0 = thin window susceptance taken from Fig. 18
λ_g = wave-guide wavelength
t = window thickness

A window of the type shown in Fig. 19 exhibits susceptance vs. frequency characteristics similar to those of a parallel resonant circuit shunting the wave guide. The resonant wavelength, at which the window introduces no reflection, may be determined approximately from the expression

$$\frac{a}{b}\sqrt{1-\left(\frac{\lambda}{2a}\right)^2} = \frac{a'}{b'}\sqrt{1-\left(\frac{\lambda}{2a'}\right)^2} \tag{26}$$

where a and b = inside guide dimensions
a' and b' = window dimensions, as shown in Fig. 19
λ = free-space wavelength

The slope at resonance of the susceptance vs. frequency curve decreases as the height of the window opening increases.

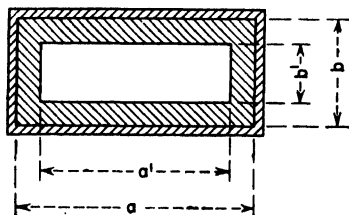



Fig. 19. Resonant window in a thin metallic diaphragm.

Metallic Posts. A round metallic post symmetrically located in a rectangular wave guide, as shown in Fig. 20, has an equivalent circuit consisting of an inductive susceptance shunted across the guide at the location of the post. The magnitude of this susceptance is given as a function of the post diameter.

If the post extends only part way across the guide, as in Fig. 21, the equivalent circuit is a capacitive susceptance, whose magnitude depends upon the length and diameter of the post. At any one frequency, as the length of the post is increased, a position is found at which the susceptance becomes infinite in magnitude. As the frequency is varied about this value, the variation of susceptance vs. frequency is

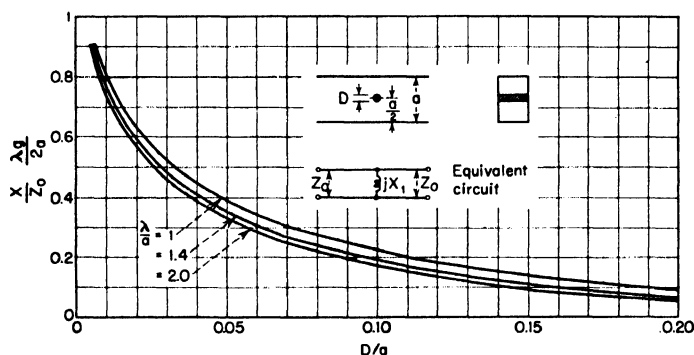



Fig. 20. Equivalent susceptance of round metallic post in a rectangular wave guide.

similar to that of a series-resonant circuit shunting the wave guide. The resonant wavelength is found when the post length is approximately one-quarter of a free-space wavelength, slightly less for posts of larger diameter.

Resonant Rings. A rectangular metal ring placed in a rectangular wave guide perpendicular to the axis also exhibits characteristics similar to a series-resonant circuit shunting the guide. The resonant wavelength depends upon the mean perimeter of the ring and is found when the perimeter is somewhat greater than a full wavelength.

Fig. 21. Resonant metallic post in rectangular wave guide.

19. Obstacles in Circular Wave Guides. For a wave guide which is above cutoff for the $TE_{1,1}$ mode only, a thin diaphragm with a centered, circular hole behaves as a shunting inductive susceptance. The magnitude of this susceptance as a function of the hole diameter is given in Fig. 22. A thin metal disk centrally located in the wave guide behaves as a capacitive susceptance shunted across the guide.

A variety of apertures in thin diaphragms act as parallel resonant circuits shunting the guide, and for each of these apertures it is usually possible to find a similar obstacle which acts as a series-resonant circuit shunting the guide. Some examples of these circuits are shown in Fig. 23.

FIG. 22. Measured equivalent susceptance of a circular hole in a thin diaphragm in a circular wave guide operating in the $TE_{1,1}$ mode. The operating wavelength was 3.2 cm; window thickness, 1/32 in.

20. Miscellaneous Wave-guide Structures.[1] *Bends.* The direction of transmission in a wave guide may be changed by bending the guide. The bent section has characteristics slightly different from the straight guide, and consequently there will be an impedance discontinuity at the junction of the bent and straight sections. A number of modes will be excited at the discontinuity and may carry away appreciable power if the guide is above cutoff for those other modes.

FIG. 23. Wave-guide structures equivalent to resonant circuits shunting the wave guide. Circular wave guide operating in the $TE_{1,1}$ mode with electric field vertically polarized.

If the guide is operating with the dominant mode and if all other modes are below cutoff, a bend is usually a very satisfactory way of changing direction of transmission. For rectangular guides, there are two types of bend, shown in Fig. 24, known as E-plane and H-plane bends, in the plane of the electric and magnetic fields, respectively. If the inner radius of the bend is a wave-guide wavelength or greater, the reflection from the bend will be small (SWR < 1.05) over the normal operating range of the wave guide. The radius may be much smaller and still yield satisfactory results if the bend is carefully fabricated. With a sharp bend, the cross section of the guide should be undistorted for best operating results.

A bend in a circular wave guide operating in the $TE_{1,1}$ mode, if not in the plane of polarization or normal to that plane, will split the transmitted wave into two components with different phase velocities and cause elliptical polarization of the trans-

[1] Moreno, T., Wave Guide Transmission Systems, *Electronics*, **19** (6), 136, June, 1946.

mitted wave. If the plane of the bend is in one of the planes of mode symmetry, the radius of the bend can be considerably less than a wave-guide wavelength, and the bend will give satisfactory results if the cross section is not distorted.

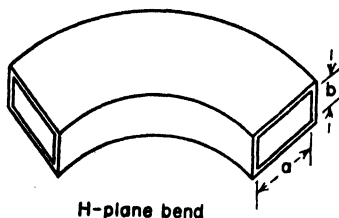
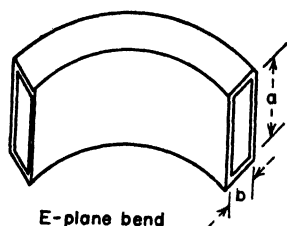



FIG. 24. Bends in rectangular wave guide.

Corners. An abrupt corner is in general not a satisfactory means of changing the direction of transmission in a wave guide, as an appreciable reflection will be set up for all but very small changes in direction. If the desired angular change is accomplished with the use of two simple corners, spaced apart approximately a quarter-wave-guide wavelength, the reflections from the two corners will tend to cancel each other. For E-plane corners, shown in Fig. 25, the optimum spacing is very close to a quarter-wave-guide wavelength at all wavelengths, but for H-plane corners, the optimum spacing is slightly different and is plotted as a function of λ/a in Fig. 25 for 90 deg total change in direction.

The band width over which these corners will introduce a standing-wave ratio less than 1.05 is between 6 and 20 per cent for a 90-deg bend, being larger for smaller values of λ/a. Circular bends are preferred for larger angles and greater band widths.

The reflection from a corner may also be eliminated by modifying the corner as shown in Fig. 26, and the band width will be comparable to that obtained with the

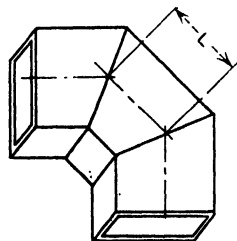
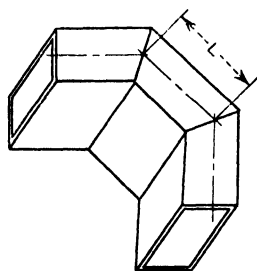
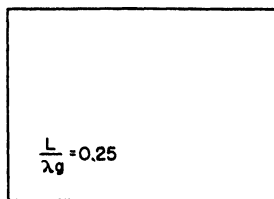
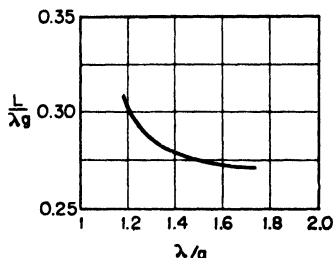



FIG. 25. Corners, 90 deg, designed for minimum reflection. Optimum dimensions for *H*-plane corners in graph; for *E* corners, in box. λ_g is wave-guide wavelength.

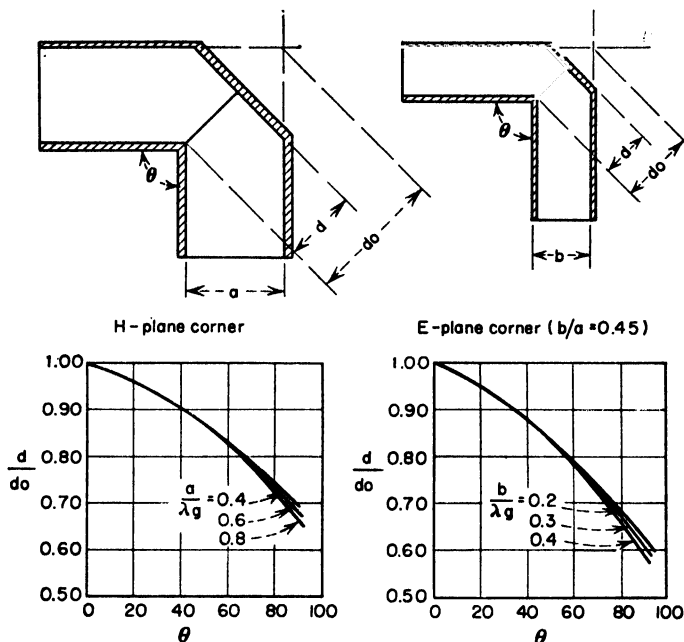



FIG. 26. Minimum reflection corners with dimensions as function of angle.

double corner. The required dimensions for E- and H-plane bends are given in Fig. 26 as a function of the angle. Manufacturing tolerances must be held rather closely on these corners for optimum performance.

Twists. The direction of polarization in a rectangular wave-guide transmission system may be changed by twisting the guide about its axis (Fig. 27). If carefully fabricated, these twists will introduce no appreciable reflection into a matched transmission system. A 90-deg twist will generally introduce a standing-wave ratio below 1.1 if more than $2\lambda_g$ in length, and shorter twists are equally satisfactory if carefully made.

FIG. 27. Twist in rectangular guide to change direction of polarization.

21. Wave-guide Connectors. Two sections of a wave guide may be connected by machining the two ends of the guide off square and clamping them together. This is a very satisfactory connection if the guides are carefully aligned and tightly clamped, and the butt joint will be electrically smooth and free from loss. The efficiency of the connection depends critically upon the accuracy of the joint, however. If the two guides are slightly misaligned or not tightly clamped, or if the ends are not finished off square and smooth, the loss and reflection may be relatively large. Losses of up to 1.0 db have been measured on butt connections that appeared satisfactory to a superficial examination. So while butt joints may be very satisfactory and are even preferred when accurate measurements are being made, they must be assembled with care and accuracy to give satisfactory performance.

The assembly of wave-guide components is facilitated with the use of choke couplings, as shown in Fig. 28. These offer advantages in flexibility and reliability over

simple butt joints because their performance does not depend critically upon the tightness of the connection. The L-shaped cavity between the choke and flange is effectively a half-wave, shorted transmission line, which is in series with the guide and offers a minimum discontinuity. The circular slot in the choke is $\lambda/4$ deep, and the actual contact between choke and flange is in series with the high-input impedance of this resonant slot, and is at a point of minimum current. For this reason, a poor contact between choke and flange does not impair the efficiency of the connection. The distance between the quarter-wave slot and the inside of the wave guide must be determined experimentally, but is effectively $\lambda/4$. For maximum band width, the slot width should be several times greater than the spacing between choke and flange faces. A properly designed choke will introduce a standing-wave ratio less than 1.05 over most of the usable range of a rectangular wave guide.

It is not necessary that the outer faces of the choke and flange be in contact for good performance. The connectors may be used as nonrigid couplings, such as are required when a shock-mounted chassis is connected to a rigidly mounted line. If a clearance of $\frac{1}{16}\lambda$ is allowed between choke and flange, a sidewise displacement of up to $\frac{1}{16}\lambda$ will not raise the loss above 0.3 db, or the standing-wave ratio above 1.3, for frequencies within a few per cent of the design frequency. A pair of chokes may be used instead of the choke-flange combination for a wobble joint and may have superior performance. With paired chokes there are often resonances encountered near the design wavelength which manifest themselves by sharp increases in the loss and standing-wave ratio.

FIG. 28. Choke-flange connector for rectangular wave guide.

22. Launching of Waves in Wave Guides. Energy traveling down a conventional transmission line may be transferred to a wave guide by terminating the transmission line in an antenna structure, which, instead of radiating energy into free space, is mounted inside the wave guide. The transition from conventional transmission line to wave guide may be made electrically smooth and reflectionless by designing the antenna so that it presents a matched load to the transmission line. By the law of reciprocity, a wave traveling from the wave guide to the transmission line will likewise not be reflected, provided the loss in the transition section is small.

The case of greatest practical interest is the launching of a $TE_{1,0}$ mode in rectangular wave guide from a coaxial transmission line. The axis of the coaxial line is generally made normal to the broad face of the rectangular wave guide. As illustrated in Fig. 29*a*, the outer conductor of the coaxial line is terminated at the wall of the wave guide, but the inner conductor extends into the wave guide, parallel to the electric lines of force, and forms a probe antenna which radiates down the guide.[1]

If the wave guide is terminated in both directions in its characteristic impedance, the power input from the coaxial line will divide equally in the two directions. To send the energy in a single direction down the wave guide, one end is shorted with a metal plate. The distance of this plate from the antenna structure and the length of the antenna are chosen to present a matched impedance to the coaxial line. Experimental determination of the optimum dimensions is usually required.

The band width over which the impedance transition between coaxial line and wave guide is relatively smooth and reflectionless can be increased by increasing the diameter of the probe antenna. Further improvements may be realized by modifying the structure as shown in Fig. 29*b*. The structures shown in Fig. 29*c* and *d* are

[1] Moreno, *loc. cit.*

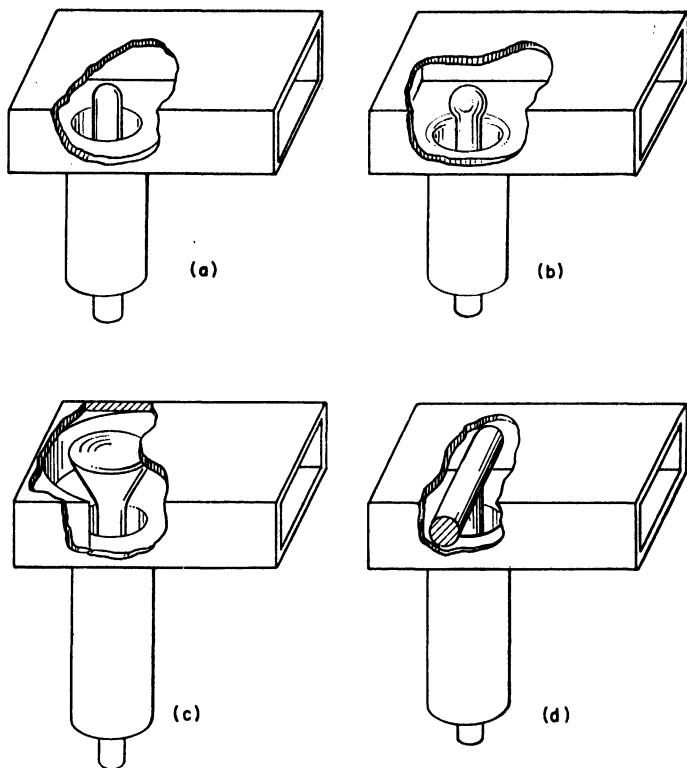



FIG. 29. Broad-band, fixed-tuned coaxial line to wave-guide transformers providing reflectionless transmission.

also broad-band transitions between coaxial line and wave guide, and provide the additional advantage of a mechanical support for the center conductor of the coaxial line. Again, the optimum dimensions for these structures must be determined experimentally.

23. Mode Transformers for Wave Guides. It is sometimes desirable in wave-guide transmission systems to change from one mode of transmission in one wave guide to a second mode of transmission in another guide. These transformers may be designed in a variety of ways, of which the following examples are illustrative:

FIG. 30. Tapered transmission between rectangular and circular wave guides, both operating in dominant mode.

An example of the simplest type of transformer is that required to change from the $TE_{1,0}$ mode in rectangular wave guide to the $TE_{1,1}$ mode in circular wave guide. Because of the similarity of these two modes, it is only necessary to construct a continuous taper in the guide cross-sectional dimensions, as shown in Fig. 30. If the taper is sufficiently long ($2\lambda_g$ is usually sufficient) the required mode transformation will be accomplished with a minimum of discontinuity and reflection in the transmission system.

A more complex transformer is required for the transition to higher modes in one

or both guides, *e.g.*, the transition from $TE_{1,0}$ mode in rectangular guide to $TM_{0,1}$ mode in a circular wave guide. Several transformers have been designed to accomplish this result, some of which are illustrated in Fig. 31. The problem is more difficult because the transition section may excite a large number of modes in the circular wave guide, and the dominant $TE_{1,1}$ mode as well as the desired $TM_{0,1}$ mode will be above cutoff. The transformer must therefore be carefully designed so as to minimize the excitation of the undesired $TE_{1,1}$ mode, whether by reasons of symmetry or by other techniques. Mode suppressors are also useful in the design of these transformers. Resonant rings are used for this purpose; when properly designed, they will completely reflect the $TE_{1,1}$ mode and have negligible effect upon the $TM_{0,1}$ mode. One of these resonant rings is shown in the transformer of Fig. 31*a*, placed in a position to minimize the excitation of the $TE_{1,1}$ mode in the circular wave guide.

FIG. 31. Transformers from $TE_{1,0}$ mode in rectangular to $TM_{0,1}$ mode in circular wave guide. In addition to providing smooth transition between desired modes, these transformers should not excite the dominant $TE_{1,1}$ mode in the circular wave guide which is above cutoff.

In Fig. 31*b*, the stub section of round wave guide extending below the junction with the rectangular guide is effectively in series with the rectangular and round guides. The length and diameter may be chosen so that the stub is effectively $\lambda/2$ long for the desired $TM_{0,1}$ mode, but effectively $3\lambda/4$ for the $TE_{1,1}$ mode. The stub therefore offers an infinite impedance in series with the guide for the $TE_{1,1}$ mode, but a zero impedance for the $TM_{0,1}$ mode. In all these transformers, a matching diaphragm may be used to eliminate any residual reflection from the transition.

24. Multimode Wave Guides. Multimode guides offer two significant advantages to the designer of microwave transmission systems. Through them, transmission over a broader band of frequencies is possible. They produce lower losses and less delay distortion than wave guides using the dominant mode. Thus, many more channels of a given width can be transmitted over a multimode guide than over a guide transmitting in the dominant mode only.

FIG. 32. Faraday rotation caused by ferrite in wave guide.

In multimode applications, each of the modes has to be launched and withdrawn without disturbing the other modes present. Recent work done on these problems will be found in the literature.[1]

[1] Beck, A. C., Measurement Techniques for Multimode Waveguide, *Trans. IRE*, **MTT-3** (3), 35, April, 1955. King, A. P., Dominant Waveguide Transmission Characteristics of a Multimode Round Waveguide, *Proc. IRE*, **40**, 966, August, 1952. Miller, S. E., and A. P. King, Low-loss Waveguide Transmission, *Proc. IRE*, **41**, 348, March, 1953. Miller, S. E., Waveguide as a Communication Medium, *Bell System Tech. J.*, **33**, 1209, November, 1954. Miller, S. E., Coupled Wave Theory and Waveguide Applications, *Bell System Tech. J.*, **33**, 661, May, 1954.

Probes, loops, and iris structures have been used successfully. Transducers located at the ends of the guide have become useful. Modes may also be separated by their discrete attenuation-coefficient values. Mode filters that use sheets of resistive material running parallel to the E vector of the mode to be attenuated have been applied successfully. It is also possible to discriminate between modes by making use of their different velocity of propagation. Multimode work permits application of real ingenuity since it is still a fast-growing field.

FIG. 33. Rotation and loss of ferrite rotator.

25. Ferrites in Wave Guides. At microwave frequencies[1] and in a static magnetic field, propagation characteristics of phase, polarization, and transmission loss vary with the magnetic field intensity in a wave guide containing ferrite materials.

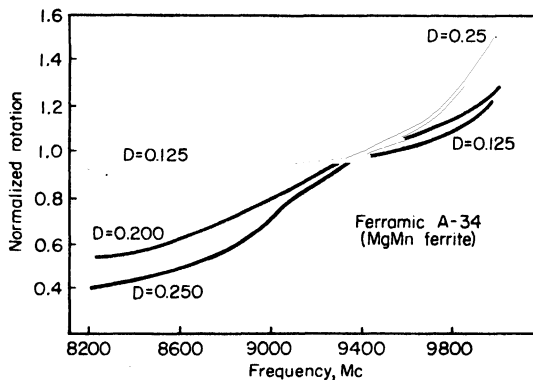



FIG. 34. Rotation as function of frequency.

Ferrite inserted into a wave guide and magnetized in the direction of propagation porduces a rotation of the plane of polarization in accordance with the Faraday effect. Figure 32 shows the rotation caused by a cylindrical piece of magnesium-manganese ferrite ¼ in. in diameter and 1½ in. long inserted into the center of a standard 1- by ½-in. wave guide at 9,375 Mc. An axial magnetic field is provided by a solenoid wound around the wave guide. At magnetic saturation, the ratio of transmission loss in the ferrite to the rotation of the plane of polarization produced by it is constant.

Figure 33 shows, at saturation, the plane-of-polarization rotation per unit length as a function of ferrite-rod diameter, and Fig. 34 shows, under saturation conditions, how normalized rotation varies with frequency for several ferrite-rod diameters.[2]

Rectangular guides may also be loaded with ferrites; Fig. 35 shows the geometry for this case.

FIG. 35. Geometry of ferrite-loaded wave guide.

[1] Hogan, C. L., The Ferromagnetic Faraday Effect at Microwave Frequencies and Its Applications, *Bell System Tech. J.*, **31**, 1, January, 1952.

[2] Sakiotis, N. G., and H. N. Chait, Properties of Ferrites in Waveguides, *Trans. IRE*, **MTT-1** (2), 11, November, 1953.

If a magnetizing field perpendicular to the broad face of the guide is applied as in Fig. 35, it is possible to obtain either reciprocal or nonreciprocal propagation through the guide. Nonreciprocal propagation is produced by nonsymmetrical insertion while reciprocal propagation is the result of symmetrical insertion; Fig. 36 shows phase shift and rotation to be identical in both directions for this case.

One-way transmission lines can be con-

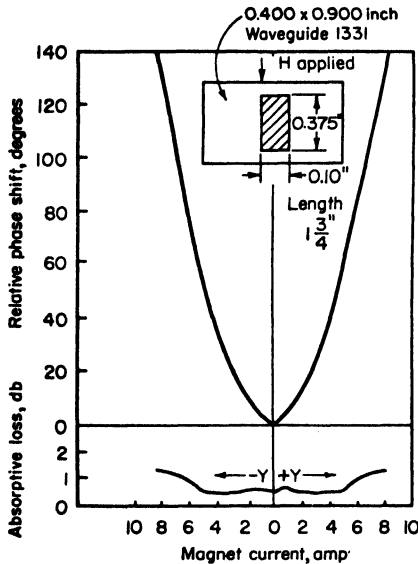



FIG. 36. Result of symmetrical placement of ferrite in wave guide.

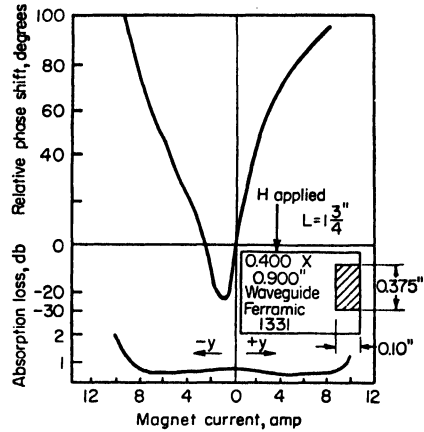



FIG. 37. Ferrite placed against one wall produces one-way transmission.

structed by placing the ferrite in the wave guide against one side wall, as in Fig. 37. This produces the desired nonreciprocal characteristics. Nonreciprocal and reciprocal behavior of ferrite-loaded wave guides is due to gyromagnetic resonance in the ferrite material. If the r-f magnetic field is circularly polarized in a plane perpendicular to the magnetization of the ferrite, the r-f power absorption in the ferrite will show a pronounced peak

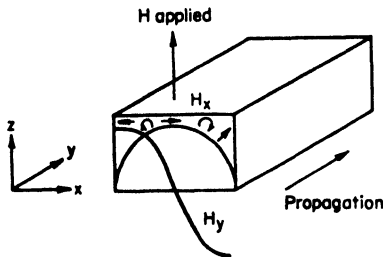



FIG. 38. Polarization of the H vector, $TE_{1,0}$ mode.

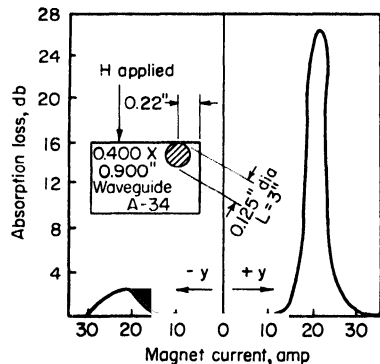



FIG. 39. One-way transmission produced by ferrite.

at a value of the magnetizing field which produces gyromagnetic resonance, provided that the sense of the circular polarization is positive relative to the magnetizing field. For a negative sense of circular polarization this absorption peak fails to materialize. These relations are shown in Fig. 38.

In wave-guide propagating in the $TE_{1,0}$ mode there are two parallel planes so

located that the H vector is polarized circularly in one sense in the one, and in the other sense in the other. The figure shows variation of the amplitudes of the two quadrature components of the H vector across the guide cross section. The two components are of equal magnitude at the two equally spaced planes mentioned above and here the magnetic r-f field is circularly polarized. The longitudinal field component, however, is of opposite sign at these planes, which indicates a 180-deg phase difference.

If some ferrite material is located at one of these planes and the static magnetic field is adjusted for gyromagnetic resonance at the r-f frequency, one direction of propagation will experience much absorption while there will be little absorption in the other. Some values measured by Sakiotis and Chait[1] are shown in Fig. 39. Application of ferrites to wave guides is still in the development stage. It has permitted many valuable systems applications.[2]

26. Coupling through Holes between Wave Guides.[3] For a rectangular wave guide operating in the $TE_{1,0}$ mode, the circuit equivalent of a small centered hole in an infinitely thin diaphragm normal to the guide axis in an inductive diaphragm normal to the guide axis is an inductive susceptance shunting the guide (Fig. 40). For waveguide as shown in Fig. 40, and having a hole of diameter d, the susceptance of the hole normalized to the guide admittance is

$$\frac{B}{Y_0} = \frac{3}{2\pi}\frac{ab\lambda_g}{d^3} \tag{27}$$

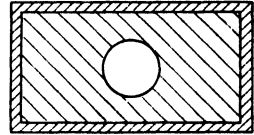



FIG. 40. Small round hole in thin diaphragm in rectangular guide is equivalent to an inductive susceptance shunting the guide.

The insertion loss of a hole in an infinitely thin diaphragm placed between a matched source and receiver is given by

$$\alpha_1 = 10 \log_{10}\left[\frac{(B/Y_0)^2}{4} - 1\right] \quad \text{db} \tag{28}$$

Finite thickness t of the diaphragm results in an additional attenuation, given by

$$\alpha_2 \cong 32.0\frac{t}{d} \quad \text{db} \tag{29}$$

The over-all attenuation is then given by $\alpha_t = \alpha_1 + \alpha_2$.

When the hole is used to couple energy between two wave guides whose axes are not coincident, the insertion loss of the coupling hole may in many cases be calculated by a simple modification of the term α_1 of Eq. (28). Examples of arrangements where such simple modifications are possible are shown in Fig. 41. The factor α_2 is not affected by the position of the hole. If the hole is in the broad face of the guides, the problem of coupling is more complex, but a complete solution has been given by Bethe.[4]

[1] *Ibid.*

[2] Cunliffe, A., R. N. Gould, and K. D. Hall, On Cavity Resonators with Non-homogeneous Media, *J. IEE*, **101**, Part III, 192, May, 1954. Van Uitert, L. G., J. P. Schafer, and C. L. Hogan, Low-loss Ferrites for Application at 4,000 Mc, *J. Applied Phys.*, **25**, 925, July, 1954. Beljers, H. G., Faraday Effect in Magnetic Materials with Traveling and Standing Waves, *Philips Research Repts.*, **9**, 131, April, 1954. Lax, B., K. J. Button, and L. M. Roch, Ferrite Phase Shifters in Rectangular Waveguide, *J. Applied Phys.*, **25**, 1413, November, 1954; New Ferrite Mode Configurations and Their Application, *J. Applied Phys.*, **126**, 1186, September, 1955. Cacheris, J., Microwave Singlesideband Modulator Using Ferrites, *Proc. IRE*, **42**, 1242, August, 1954. Berk, A. D., and B. A. Lengyel, Magnetic Fields in Small Ferrite Bodies with Applications to Microwave Cavities Containing Such Bodies, *Proc. IRE*, **43**, 1587, November, 1955. Rich, J. A., and S. E. Webber, Ferrite Attenuators in Helixes, *Proc. IRE*, **43**, 100, January, 1955. Gunn, J. B., and C. A. Gogarth, A Novel Microwave Attenuator Using Germanium, *J. Applied Phys.*, **26**, 353, March, 1955. Damon, R. W., Magnetically Controlled Microwave Directional Coupler, *J. Applied Phys.*, **26**, 1282, October, 1955. Sullivan, R. F., and R. C. LeCraw, New Type Ferrite Microwave Switch, *J. Applied Phys.*, **26**, 1282, October, 1955.

[3] Bethe, H. A., Theory of Diffraction by Small Holes, *Phys. Rev.*, **66** (7, 8), Oct. 1 and 15, 1944.

[4] *Ibid.*

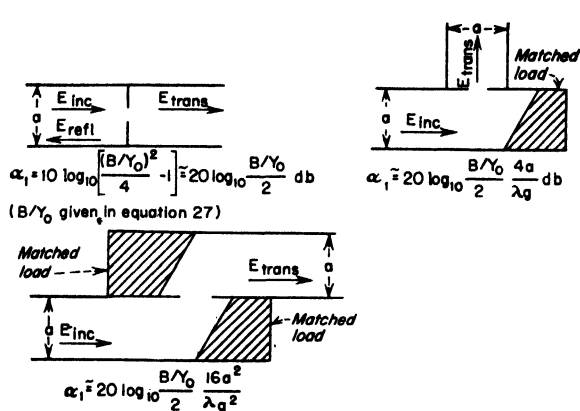



FIG. 41. Attenuation through small hole separating two sections of rectangular guide, under various conditions of operation.

27. Wave-guide Tees and Junctions.[1] The joints in wave guides are of two types, as shown in Fig. 42.

1. *H-plane Tee or Shunt Tee.* The branch guide is taken off the narrow side of the main guide in the plane of the magnetic field. This is called a *shunt tee* because a signal fed in the branch guide will divide at the junction and be in phase in arms 1 and 3 at points equidistant from the junction. The equivalent circuit for two-wire transmission lines is shown in Fig. 42.

2. *E-plane or Series Tee.* The branch guide is taken off the wide side of the main guide in the plane of the electric field. This is called a *series tee* because a signal fed in the branch guide will divide at the junction and be out of phase in arms 1 and 3 at points equidistant from the junction. The equivalent circuit for two-wire transmission lines is shown in Fig. 42.

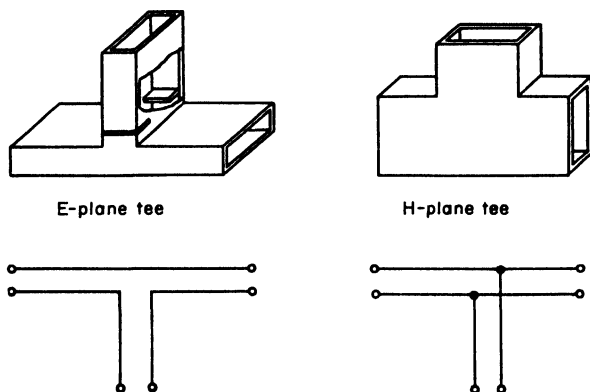



FIG. 42. Two kinds of tee joints in rectangular wave guide and their two-wire transmission-line equivalent circuits. The E-plane tee is partially cut away to show an added matching diaphragm.

As would be expected from the equivalent circuits for these tee joints, positions may be found in the branch arms at which short circuits may be placed to give reflectionless transmission past the branch, and other positions may be found at which a short circuit may be placed to give a complete reflection of energy at the branch arm. But because of the distortion and fringing fields at the junction, the positions at which these short circuits must be placed to realize these effects will be somewhat different from the values predicted by conventional transmission-line theory, which neglects these field distortions. For example, with the shunting arm, one would expect that a short circuit should be placed $\lambda/4$ or $3\lambda/4$ away from the junction to give reflectionless transmission past the branch arm. Experimentally, this value is found to vary

[1] Allanson, J. T., R. Cooper, and T. G. Cowling, The Theory and Experimental Behavior of Right-angled Junctions in Rectangular-section Wave Guides, *J. IEE*, **93** (23), Pt. III, 177.

with ratio of operating wavelength to cutoff wavelength, a typical mid-band value being 0.57λ from short to inside face of the main guide.

Judging by the equivalent circuit, it should be possible to obtain reflectionless transmission around a corner by proper location of a short in the straight-through arm of the wave guide. Experimentally it is found that the reflection may be minimized but not eliminated by proper positioning of the short. This minimum reflection varies with operating wavelength; in typical operation the minimum standing-wave ratio may be approximately 1.3.

To obtain reflectionless transmission around a corner with a tee junction, it is necessary to add some sort of matching structure, such as a diaphragm, the reflection from which is adjusted to cancel the residual reflection from the tee. A typical matching structure of this sort as placed in the branch arm is shown in Fig. 42.

Fig. 43. Magic tee in rectangular guide.

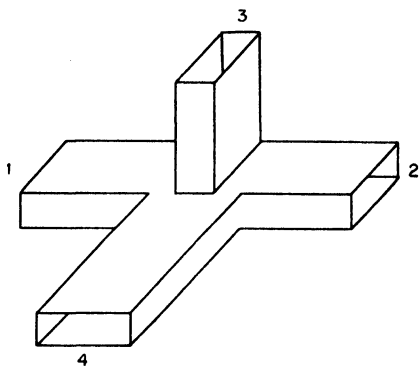
28. Wave-guide Bridge Structures. A variety of wave-guide structures may be built which have properties similar in many respects to l-f bridge circuits. These structures utilize tee joints of the two types discussed in the previous paragraphs.

The best known of these bridge circuits is the *magic tee*, or *hybrid tee*, shown in Fig. 43. Two branch wave guides are connected to a colinear wave guide at the same point, one in an H plane, shunt tee and the other in an E plane, series tee. If a signal is fed in the shunt arm 4, and the two colinear arms 1 and 2 are terminated in their characteristic impedances, the input power will, for reasons of symmetry, divide equally between the two load impedances. Also for reasons of symmetry, there will be no signal delivered into arm 3, as there is no net electric vector developed across the entrance to this series arm. The input impedance to the shunt arm will not in general match the characteristic impedance of the wave guide; *i.e.*, there will be standing waves in the input arm.

Fig. 44. Symmetrical four-terminal equivalent of magic tee.

A similar situation will exist if power is fed into the series arm. If the colinear arms 2 and 1 are terminated in their characteristic impedances, the input power will divide equally between these two arms, and no power will be delivered to the shunt branch arm 4. There will be standing waves set up in the input arm, however.

The symmetry of the structure will not be destroyed if the two matched loads on the colinear arms are replaced by any two impedances that are equal in magnitude and phase angle, and under these conditions there will still be no coupling between the series and the shunt branch arms. It is apparent that the behavior of the structure is similar to a bridge circuit, with the signal fed in one branch arm and the null detector on the other.

With matched loads on the two colinear arms, the impedance measured looking into either of the branch guides does not match the guide impedance. If, by the addition of some sort of matching elements, these impedances are made to match the wave-guide impedance, so that there are no standing waves in the input branch arms, the structure will possess additional qualities of balance.

As before, if power is fed in either branch arm and if the two colinear arms are terminated in matched impedances, the input power will divide equally between these impedances, and there will be no coupling between the two branch arms. Because of the added matching structures, a matched impedance will be seen at the input. In addition, if a signal is fed in one of the colinear arms, it will divide equally between the two branch arms, if they are matched, and there will be no coupling to the other colinear arm. Also, the input impedance to the colinear arms under these conditions will match the guide impedance.

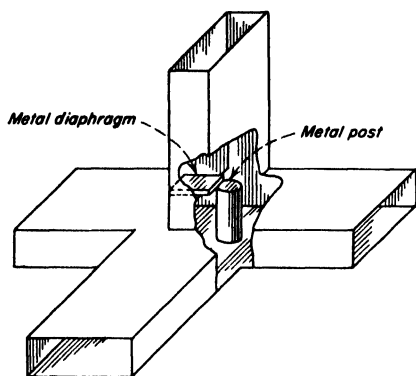



FIG. 45. Cutaway view of magic tee showing impedance-matching diaphragm and post.

FIG. 46. Hybrid coil equivalent of magic tee of Fig. 43.

The matched structure is equivalent to a balanced four-terminal-pair network, shown schematically in Fig. 44, in which each wave guide is represented by a terminal pair. In this equivalent network, if a signal is fed into any terminal pair and if the two adjacent terminal pairs are terminated in matched loads, the power will divide equally between these loads, and no voltage will appear across the opposite terminal pair.

The matching of the branch arms may be accomplished in a variety of ways. For maximum band width, it is desirable to have the matching structure as close as possible to the junction. One typical design is shown in Fig. 45, in which the H-plane tee is matched by the post and the E-plane tee by the asymmetrical inductive diaphragm.

The l-f equivalent circuit of the magic tee is the hybrid coil, shown in Fig. 46, and the properties possessed by the two circuits are equivalent in many ways.

FIG. 47. Rat-race wave-guide bridge, electrically equivalent to magic tee of Fig. 43. The four branching guides connect with the ring in series tees. Distances between branch guides are measured along the mean circumference of the ring.

Several other structures possess properties electrically equivalent to the magic tee. One of these, known as a "ring" or "rat-race" bridge, is shown in Fig. 47. The distance between any two of the branch arms may be increased by an integral multiple number of half wavelengths without affecting the properties of this circuit. For reasons of impedance matching, it is desirable that the height of the wave guide forming the ring be somewhat less than the height of the four incoming guides. The band width of this structure is less than that of the magic tee, but the power-handling capacity is greater.

MISCELLANEOUS WAVE-GUIDE DEVICES

29. Standing-wave Detectors (Impedance Meters). To measure the standing waves in a wave guide, the field strength in the guide must be measured as a function of position along the guide. The fields are completely contained within the outer conducting tube, so this tube must be slotted to allow entry of the probe or loop that samples the field. The slot should have a minimum effect upon the fields in the guide and should therefore be located where there is no current in the wave-guide walls transverse to the slot. For a rectangular wave guide operating in the $TE_{1,0}$ mode, the preferred location is in the center of the broad face of the guide, where the current has only an axial component (Fig. 48).

FIG. 48. Slot for measuring standing waves.

A small fraction of the energy contained in the guide is coupled to a detector by a small probe or loop which extends into the guide through the slot. The coupling probe or loop is mounted on a carriage which moves along the guide, and the signal coupled to the detector depends upon the field strength at the location of the coupling element.

A coupling probe is usually preferred to a coupling loop because of the probe's mechanical simplicity, and because its position may be precisely determined. The probe should not extend into the guide far enough to disturb the field pattern inside the guide, and a penetration of 15 per cent or less of the guide height is considered good operating practice.[1] For this reason, accurate mechanical construction of standing-wave detectors is essential to operating accuracy, for slight variations in probe penetration will cause corresponding variations in the indicated field strength.]

30. Matched Terminations for Wave Guide. Terminations for wave guides which present an impedance that matches the wave-guide impedance may be constructed in a variety of ways. The most successful designs, which remain matched over a considerable band of frequencies, utilize a section of wave guide to which some lossy or resistive element has been added to make it highly attenuating, with a tapered transition between the attenuating section and the incoming wave guide.

FIG. 49. Tapered strip of resistance film mounted inside rectangular guide to provide matched termination.

A relatively lossy dielectric filling the wave guide makes it highly attenuating. The phenolic resin-bonded plastics are satisfactory for this purpose, as are asbestos boards such as Transite which will handle higher powers. The tapered section may be constructed by tapering the dielectric; a taper of the type shown in Fig. 15 has proved satisfactory. If this taper is longer than 2λ, the design is very uncritical to either taper length or dielectric properties, provided a sufficient length of the high-loss dielectric is used in the attenuating section.

For lower power work, resistive films on a dielectric backing are very satisfactory. A platinized film on glass has been used, and carbon films on plastic board are very easy to work with. For maximum attenuation, the resistive film should lie in the plane of the electric field, in a region of strong field. For a rectangular guide in the $TE_{1,0}$ mode, the attenuation is high if the resistive film is placed parallel to the narrow faces of the guide and centered with respect to the broad faces. The attenuation of this lossy section is matched to the input guide by tapering the film to a point as shown in Fig. 49.

[1] Altar, W., F. B. Marshall, L. P. Hunter, Probe Error in Standing Wave Detectors, *Proc. IRE*, **34** (1), 33P, January, 1946.

31. Fixed Attenuators.[1] Fixed attenuators for wave guides can be constructed in a manner similar to matched terminations. The amount of attenuation may be controlled by varying the length of the lossy section, as well as the attenuation per unit length. For fixed attenuators, it is usually necessary to match both ends of the lossy section to the wave guide, so there are tapers at each end of the lossy section. The tapers may be replaced by steps, similar to $\lambda/4$ matching sections. These are physically shorter but are matched over a reduced band width. The attenuation resulting from the use of lossy dielectrics is usually a function of the temperature of the dielectric, and also varies with its exposure to humidity. The resistive films are more satisfactory in this respect. A typical fixed attenuator using a resistive film is shown in Fig. 50.

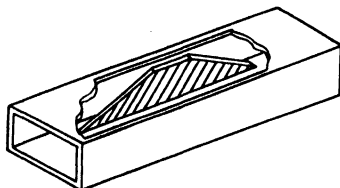



Fig. 50. Tapered resistance film to form fixed attenuator inside rectangular guide.

Directional couplers, discussed in a subsequent paragraph, are among the most satisfactory fixed attenuators.

32. Variable Attenuators. Variable attenuators used with wave guides fall into two classes: those utilizing the resistive attenuation of lossy dielectrics or resistance films, and those utilizing the reactive attenuation of below-cutoff wave guides. A typical resistive attenuator uses a resistive film of the same type used in matched terminations or fixed attenuators. The attenuation is varied by changing the position of the film within a wave guide, or by varying the amount of resistive film inserted into the wave guide through a slot in its outer wall. An example of the latter type is shown in Fig. 51. These attenuators can be accurate and stable, if carefully designed, and have the distinct advantage that the minimum attenuation may be reduced essentially to zero by removing the film completely from the guide or by moving it to a region of negligible field strength. However, these attenuators require calibration against some standard, at the same or at a different frequency.

The reactive attenuation in a below-cutoff wave guide may be used in a variable attenuator,[2] the variation being accomplished by changing the length of the below-cutoff wave guide. The attenuation in a below-cutoff wave guide may be calculated from the guide dimensions. For the $TE_{1,1}$ and $TM_{0,1}$ modes in round wave guide, the attenuation is given in Eqs. (16) and (17).

Fig. 51. Use of flap of resistance material as a variable attenuator by lowering it into a region of strong electric field.

Typical attenuator designs, in which the below-cutoff wave guides are fed from coaxial lines, are shown in Fig. 52. The input impedance to a below-cutoff wave guide is a reactance, so if the input and output impedances of the attenuator are to match the line impedances, some lossy elements must be added. These may be simply added lengths of lossy lines, or they may be shunt resistance films, as shown schematically in Fig. 52.

For the attenuation in the below-cutoff guides to equal the values given by the above formulas, the launching mechanism should excite only the desired mode in the below-cutoff wave guide. The attenuation in circular wave guide is lowest with the $TE_{1,1}$ mode, so if other modes are excited by the launching mechanism, they will attenuate more rapidly with distance, and the rate of attenuation will approach the

[1] Gaffney, F. J., Microwave Measurements and Test Equipments, *Proc. IRE*, **34** (10), 775, October, 1946.

[2] An example is given in H. R. Meahl, An Attenuator of S-band Energy, *Proc. IRE*, **35** (2), 211, February, 1947.

theoretical value of the $TE_{1,1}$ mode at large attenuations. But attenuators using the $TM_{0,1}$ mode must be very carefully designed so that the undesired $TE_{1,1}$ mode will not be excited; otherwise the $TE_{1,1}$ mode will become increasingly strong relative to the desired $TM_{0,1}$ mode at large spacings and cause a departure from linearity.

Below-cutoff wave-guide attenuators are widely used where a precision variable attenuator is desired, as in signal generators, but their primary disadvantage is the

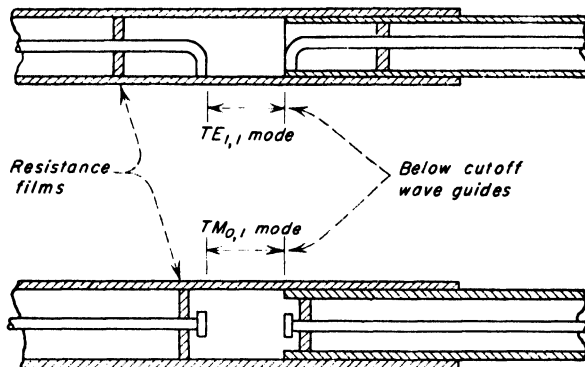



FIG. 52. Variation in below-cutoff sections as variable attenuators. Resistance films match the below-cutoff section to the input coaxial lines.

large minimum attenuation. If the input and output coupling elements are brought too close together, the calibration will deviate greatly from theoretical values, and in addition the interaction between input and output will vary the impedances presented by the attenuator to the incoming line.

33. Impedance Transformers. Fixed impedance-transforming elements in wave guides frequently consist of the metallic diaphragms described in Sec. 18. These are equivalent to susceptances shunting the wave guide and are employed in the same fashion as stub transmission lines are used to tune conventional transmission-line systems; *i.e.*, if their size and position are correctly determined, these diaphragms are capable of matching any load impedance not a pure reactance to the characteristic impedance of a wave guide.

For tunable impedance-matching devices in wave guides, shorted stub lines of variable length may be used in various combinations, and the series or shunting reactance of the stub guides is varied by changing their lengths. Another type of variable reactance that is well suited to wave-guide structures is the adjustable probe, shown in Fig. 53. The shunt reactance of this probe is varied by changing its penetration into the guide. A typical impedance transformer may utilize two or three of these adjustable probes, fixed in position and spaced along the line. Another very useful design has the probe mounted on a sliding carriage and extending through a slot into the wave guide. With this arrangement, both the penetration and position of the probe may be continuously varied.

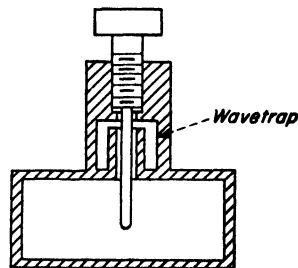



FIG. 53. Adjustable probe providing variable reactance shunting the guide. Wave trap gives electrical contact between probe and wave guide without metal-to-metal contact at inner surface of guide.

A special type of impedance transformer is the phase shifter, which is a section of a wave guide whose electrical length is continuously variable. A typical phase shifter, shown in Fig. 54, uses a tapered block of dielectric which partly fills the wave guide. The effect of the dielectric in a wave guide is to reduce the wave-guide wave-

length, but the amount of reduction depends upon the electric field strength where the dielectric is located. So in the design of Fig. 54, as the dielectric block is moved toward the center of the wave guide to a region of greater electric field strength, the wavelength in that section of wave guide will be reduced, and the electrical length of the section of guide will correspondingly increase.

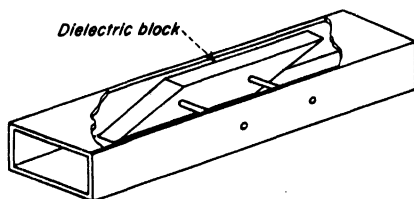



FIG. 54. Continuously variable phase shifter. The electrical length of the wave-guide section is increased by shifting the tapered dielectric block toward center of guide to region of greater electrical field.

34. Directional Couplers.[1] Directional couplers are measurement devices used in conjunction with h-f transmission lines, such as wave guides. They consist of two transmission lines, coupled together at an intersecting point or in an intersecting region. The coupling is in a manner such that an electromagnetic wave traveling in a single direction along one line will induce a wave traveling in a single direction along the second line. A wave traveling in the opposite direction on the second line will be induced by a wave traveling in the opposite direction along the first line.

The performance of a directional coupler may be described in terms of two quantities:

1. *Attenuation.* The attenuation of a coupler is the relative strength of the traveling wave on one line and the traveling wave which it induces on the second.

2. *Directivity.* Only in the ideal coupler does a single traveling wave on one line induce a wave traveling in a single direction on the second line. In a practical coupler, two waves traveling in opposite directions are induced. These are normally greatly unequal in strength, and their relative strength is the directivity of the coupler.

Directional couplers are widely used as fixed attenuators because the attenuation depends upon the coupling between lines. This coupling is generally through holes in the wall separating the two guides and is therefore quite unaffected by changes in temperature and humidity. Directional couplers also have application in the measurement of reflection coefficients and, in addition, permit a signal to be introduced in a single direction on a transmission line without actually breaking into the line.

The schematic diagram of a typical directional coupler is shown in Fig. 55. The two transmission lines are weakly coupled together by two coupling links, spaced $\lambda/4$, as shown. A wave traveling from left to right on the primary line induces through each coupling link two waves traveling in opposite directions on the secondary line. Because of the spacing between coupling links in the primary line, link B will be excited 90 deg in phase behind link A. Because of the spacing between coupling links in the secondary line, the signal traveling to the right from link A will be retarded 90 deg with respect to the corresponding signal from link B, and the two waves will therefore reinforce each other. But the signal traveling to the left from link B will be retarded an additional 90 deg with respect to the corresponding signal from link A; the two

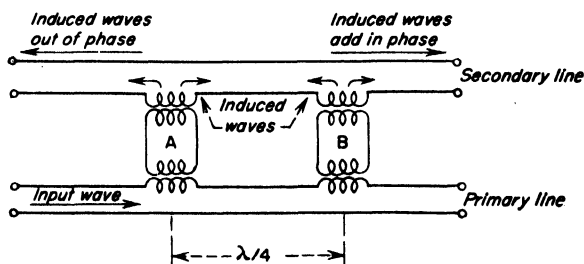



FIG. 55. Two-element directional coupler.

[1] Mumford, W. W., Directional Couplers, *Proc. IRE*, **35** (2), 160, February, 1947. Early, H. C., A Wide-band Directional Coupler for Wave Guide, *Proc. IRE*, **34** (11), 883, November, 1946.

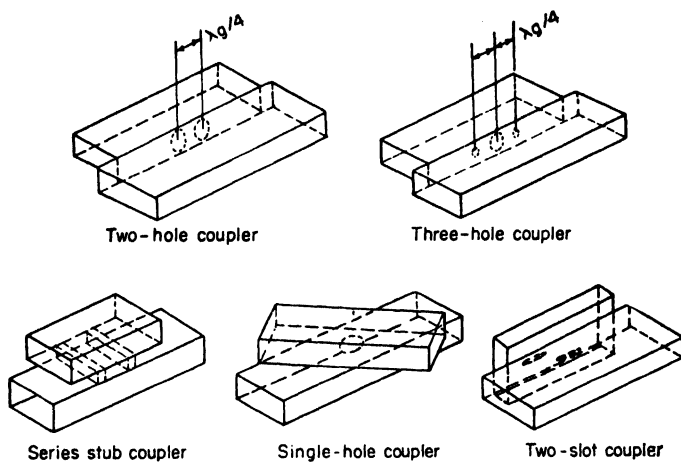



Fig. 56. Typical directional couplers for rectangular wave guide.

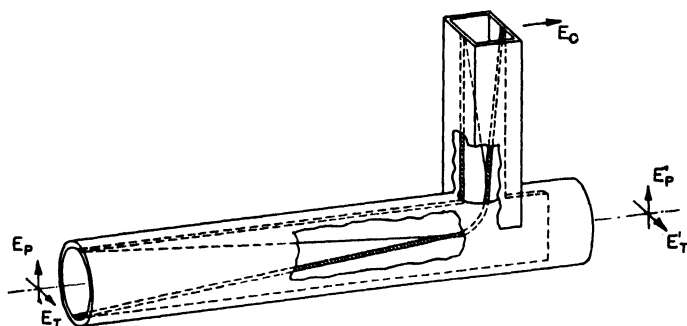



Fig. 57. Finline coupler.

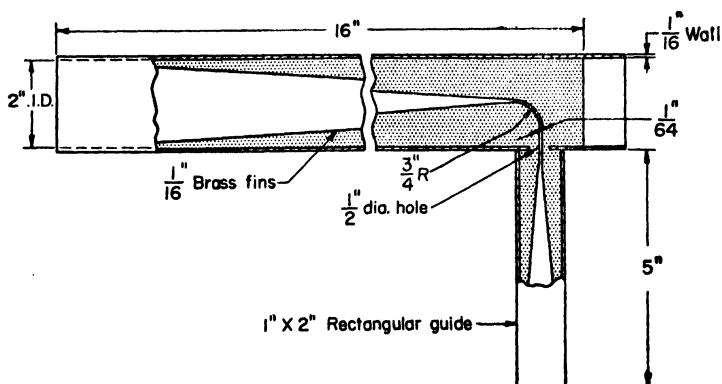



Fig. 58. Coupler construction.

signals will therefore be 180 deg out of phase and will cancel each other. As a result of these actions, the wave traveling from left to right on the primary line induces on the secondary line only a wave traveling from left to right. Correspondingly, a wave traveling from right to left on the secondary line is induced by a wave traveling in that direction on the primary line.

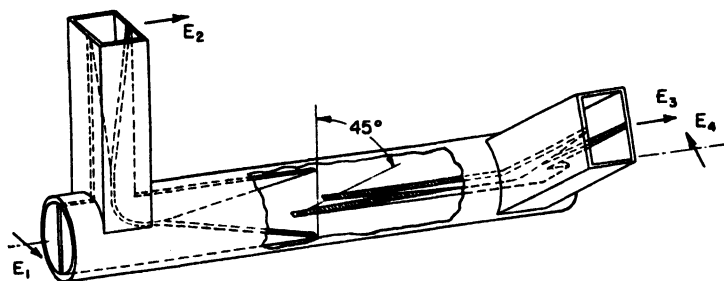



FIG. 59. Adjustable finline double coupler.

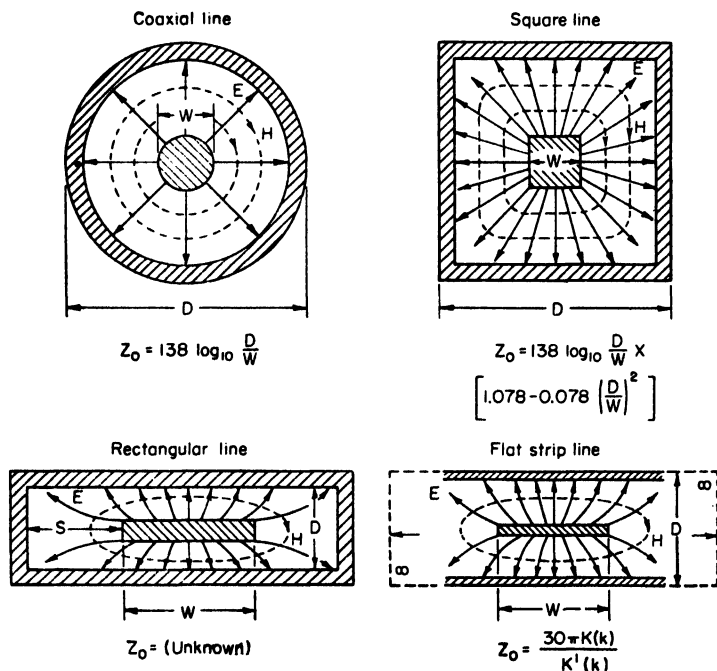



FIG. 60. Evolution of flat-strip transmission line.

The coupling elements usually take the form of apertures through the walls of the adjacent wave guides, through which energy is coupled. A number of typical directional couplers are shown in Fig. 56. The two-hole coupler is the wave-guide analogue of the two-element coupler whose operation is described above. The three-hole coupler operates in a similar manner but has high directivity over a greater band of frequencies. The series-stub coupler is similar, but the coupling elements are sections of low-impedance wave guide. The single-hole coupler obtains its directive effect

by having both electric and magnetic coupling through the hole, which reinforce in one direction and cancel in the other. The angle between the guides is chosen so that the two signals are equal in magnitude. The two-slot coupler has high directivity over a broad band because the 180-deg phase shift between the two coupled signals required for cancellation in one direction is accomplished by coupling to elements of magnetic field in one guide that are 180 deg out of phase, this phase difference being independent of frequency.

35. Finline Couplers.[1] The finline coupler permits successful operation over wide

FIG. 61. Fields in flat-strip transmission line.

[1] Robertson, S. D., The Ultra-bandwidth Finline Coupler, *Proc. IRE*, **43**, 739, June, 1955; also in *Trans. IRE*, **MTT-3** (6), 45, December, 1955. For recent work on directional couplers see Cook, J. S., Tapered Velocity Couplers, *Bell System Tech. J.*, **34**, 807, July, 1955; Fox, A. G., Wave Coupling by Warped Normal Modes, *Bell System Tech. J.*, **34**, 823, July, 1955; Monteith, G. D., Coupled Transmission Lines as Symmetrical Directional Couplers, *J. IEE*, **102**, Part B, 383, May, 1955; and Andrews, P., A Simple Waveguide Directional Coupler, *J. British IRE*, **15**, 112, February, 1955.

FIG. 62. Characteristic impedance of flat stripline. Zero-thickness case.

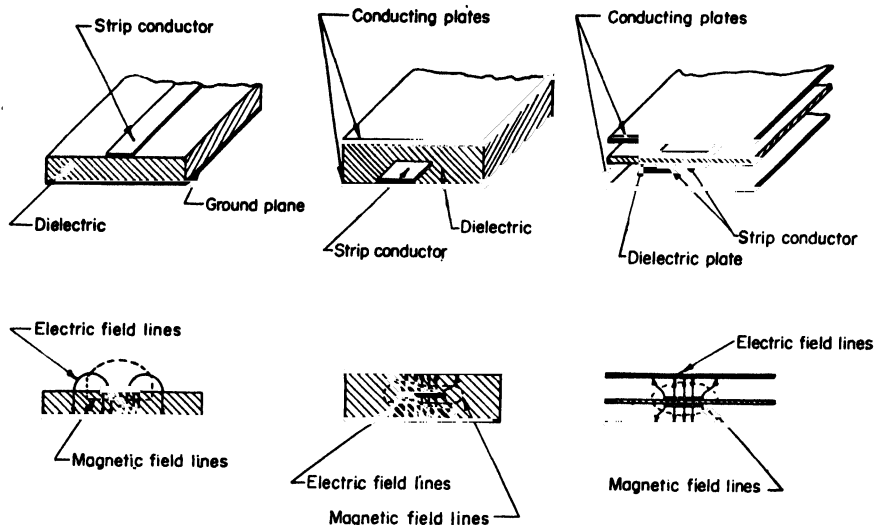



FIG. 63. Microwave striplines. (*Left*) Microstrip. (*Center*) Sandwich line. (*Right*) Strip-line.

frequency bands. It serves to build up directional couplers, hybrid junctions, and polarization-sensitive devices capable of operating over one or several octaves. The basic device is shown in Fig. 57. A rectangular wave guide is joined to a circular one. Thin metallic fins are inserted diametrically opposed to each other as in Fig. 58. Almost all the energy associated with the field whose electrical vector is parallel to the fins E_P will be guided by the fins and coupled through the hole into the rectangular guide. Energy associated with a transverse field E_T is hardly influenced by the presence of the fins and passes through the coupler. Two waves polarized in two perpendicular planes can be readily separated.

FIG. 64. Distribution of power flow for wide strip of zero thickness above infinite ground plane.

To vary the degree of coupling, two finline couplers may be combined as shown in Fig. 59. The angle between the two rectangular guides determines the distribution of energy between E_3 and E_4. At 45 deg the power is evenly split. Using E_1 as input, the energy is coupled to E_3 and E_4. If the input is at E_4, the energy is divided between E_1 and E_2.

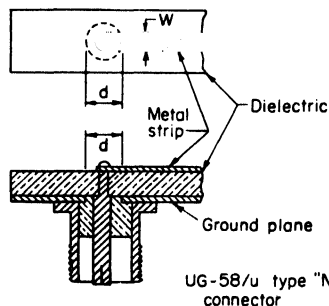



FIG. 65. Microstrip to coaxial transducer.

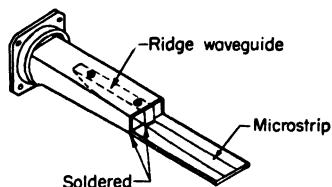



FIG. 66. Wave guide to microstrip transducer.

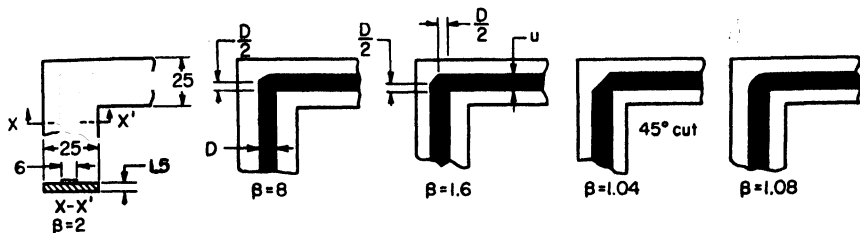



Fig. 67. Right-angle bends in microstrip. Dimensions in millimeters. Frequency, 4,700 Mc. Fiberglas G-6 dielectric. β-input VSWR.

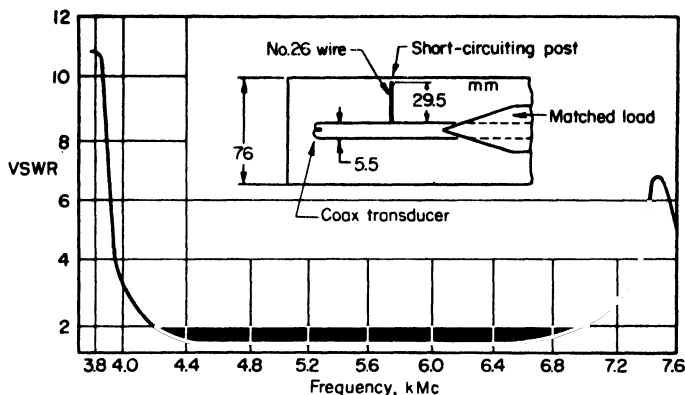



Fig. 68. VSWR vs. frequency for short-circuited stub across microstrip. Teflon-impregnated Fiberglas dielectric.

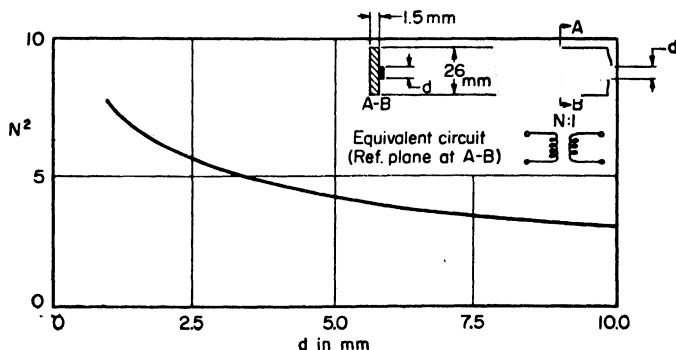



Fig. 69. Microstrip step discontinuity. Dielectric Fiberglas G-6 dielectric, frequency 4,700 Mc.

36. Planar or Flat-strip Transmission System.[1] This was probably first used by V. H. Rumsey and H. W. Jamieson at the U.S. Naval Research Laboratory during the Second World War. Since 1949 much theoretical and practical work has been

[1] Begovich, N. A., and A. R. Margolin, Theoretical and Practical Studies of a Strip Transmission Line, *Internal Technical Memorandum*, 234, Hughes Aircraft Company. Barrett, R. M., and M. H. Barnes, Microwave Printed Circuits, National Conference on Airborne Electronics, IRE, Dayton, Ohio, 1951; *Proc. IRE*, MTT-3 (2), March, 1955.

carried out in its direction. In October, 1954, a symposium on microwave strip circuits was held at Tufts College, Medford, Mass., at which a large number of developments were reported.[1]

The flat-strip system has developed rapidly. Printed-circuit techniques have been applied successfully, and today it begins to rival the hollow-metal-tube system.[2]

The planar-transmission system constitutes an evolution of the coaxial system and operates in the *TEM* mode. This evolution is shown in Fig. 60, and field distribution

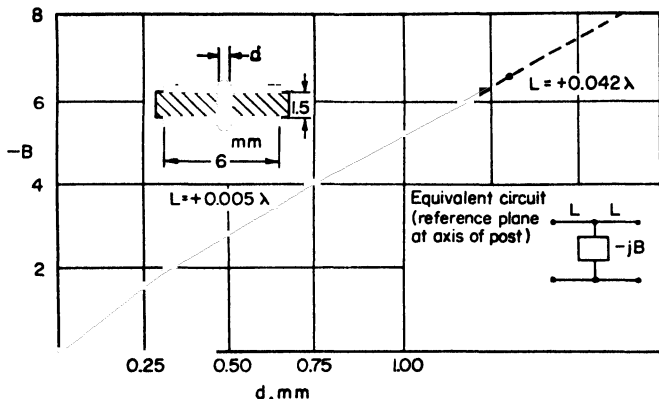



Fig. 70. Variation of shunt susceptance with diameter of post. Fiberglas G-6 dielectric, frequency 4,700 Mc.

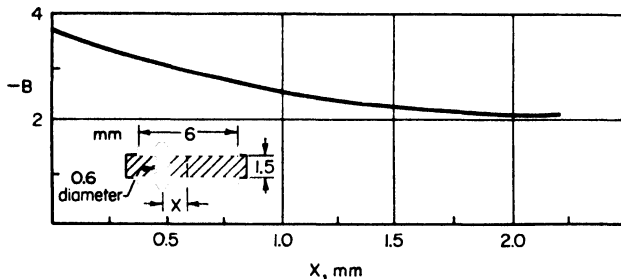



Fig. 71. Variation of shunt susceptance with location of post.

is shown in Fig. 61.[3] Physically wide transmission lines of low characteristic impedance may be analyzed using the parallel-plate capacitance formula. For lines of 50 ohms and up, fringing effects play an appreciable role and Fig. 62 shows this. It is plotted for center conductors of zero thickness. This type of line is well adapted to printing techniques, and a complete line of circuit elements and components has been developed.

[1] Special Issue, *Trans. IRE*, **MTT-3** (2), March, 1955, 21 papers.

[2] Park, D., Planar Transmission Lines, *Trans. IRE*, **MTT-3**, 8, April, 1955; *Trans. IRE*, **MTT-3, 7**, October, 1955. Dahlman, B. A., A Double-ground-plane Strip-line System for Microwaves, *J. IEE*, **102**, Part B, 488, July, 1955; also *Trans. IRE*, **MTT-3**, 52, October, 1955. Hayt, W. H., Jr., Potential Solution of a Homogeneous Stripline of Finite Width, *Trans. IRE*, **MTT-3**, 16, July, 1955. Burshtein, E., and L. Solov'ev, Propagation of a Fundamental Wave between Parallel Surfaces, *Compt. rend. acad. sci.* (*URSS*), **101**, 465, Mar. 21, 1955 (in Russian). Arditi, M., and J. Elefant, Microstrip Applied to Band-pass Microwave Filters, *Elec. Commun.*, **32**, 52, March, 1955. Cohn, S. B., Shielded Coupled-strip Transmission Line, *Trans. IRE*, **MTT-3**, 29, October, 1955. Bradley, E. H., Design and Development of Strip-line Filters, *Trans. IRE*, **MTT-4** (2), 86, April, 1956.

[3] Barrett, M., *loc. cit.*

Various groups have shared in the development of the planar-transmission system. Figure 63 shows three practically useful configurations.[1] Microstrip, shown on the left, has been used to realize a number of complex microwave circuits. It shows considerable dielectric losses and is thus not eminently suitable for long-line or high-Q filter applications. An analysis of the system is given by Deschamps.[2]

The sandwich and particularly the stripline, shown in the center and to the right in Fig. 63, permit construction of high-Q components since losses are mostly in the conductors. They are naturally more difficult to build.

In microstrip, *TEM*-mode propagation may only be assumed in a lossless system with dielectric of infinite width. Figure 64 shows, however, that field and power flow are in the dielectric near the narrow strip.

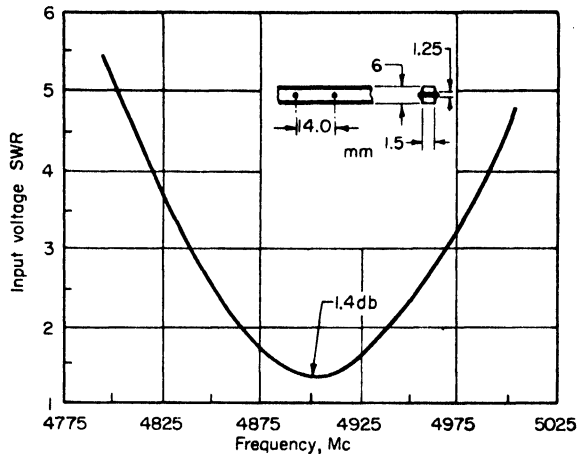



FIG. 72. Microstrip resonant section with transverse posts. Fiberglas dielectric. Insertion loss 1.4 db at center frequency.

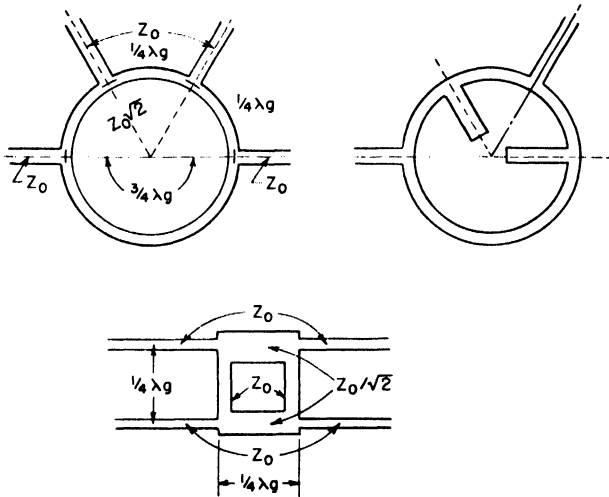



FIG. 73. Microstrip hybrid rings.

Components for use with microstrip have been developed, and several are shown in Figs. 65 to 74 from Arditi.

[1] Arditi, M., Characteristics and Applications of Microstrip for Microwave Wiring, *Trans. IRE*, **MTT-3** (2), 31, March, 1955.

[2] Deschamps, G. W., Theoretical Aspects of Microstrip Waveguides, *Trans. IRE*, **MTT-2** (1), 100, April, 1954.

37. Dielectric-image Lines. The millimeter-wave region is attracting much interest. Hollow metal wave guides with their dimensional correspondence to wavelength require an uncomfortable degree of precision in this range. Low-loss dielectric guides carry only a small fraction of the energy associated with the wave in the dielectric. The balance propagates in the regions around the dielectric. Although these regions measure several wavelengths the millimeter region permits physically realizable structures of this type.

In the dielectric, an HE_{11} or dipole mode offers low attenuation and a high degree of freedom from higher modes. Its phase constant is very nearly that of free space.[1] This seems reasonable since most of the energy travels through the space adjacent to the guide. This dipole mode possesses a plane of symmetry and may therefore

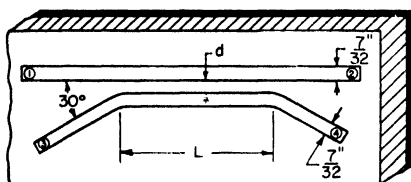



FIG. 74. Microstrip directional coupler.

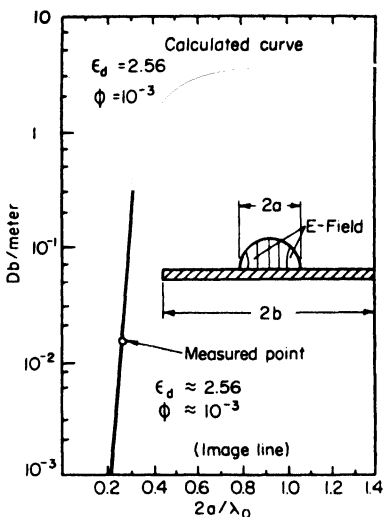



FIG. 75. Image-line cross section and attenuation curve with measured point.

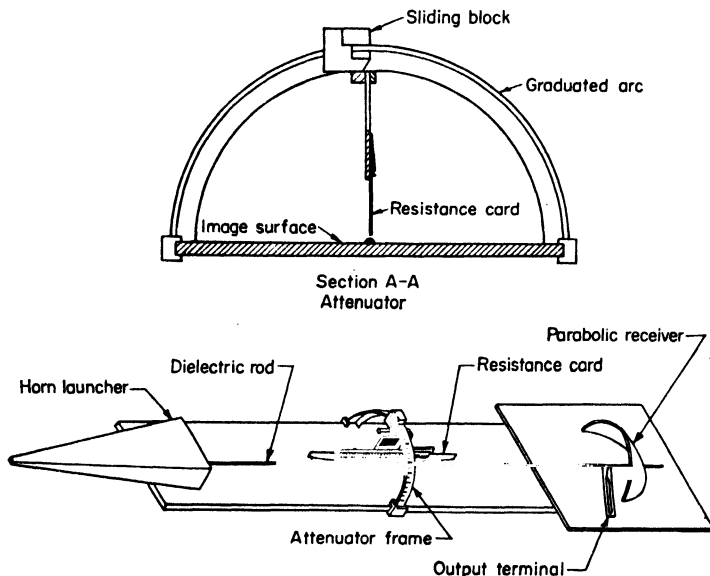



FIG. 76. Adjustable attenuator.

[1] Elsasser, W. M., Attenuation in Dielectric Circular Rod, *J. Applied Phys.*, **20**, 1192, December, 1949.

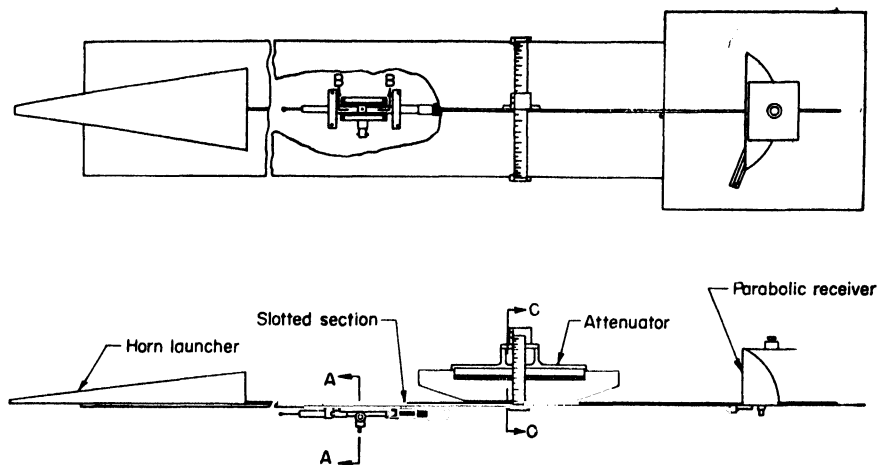



FIG. 77. Standing-wave-measuring apparatus.

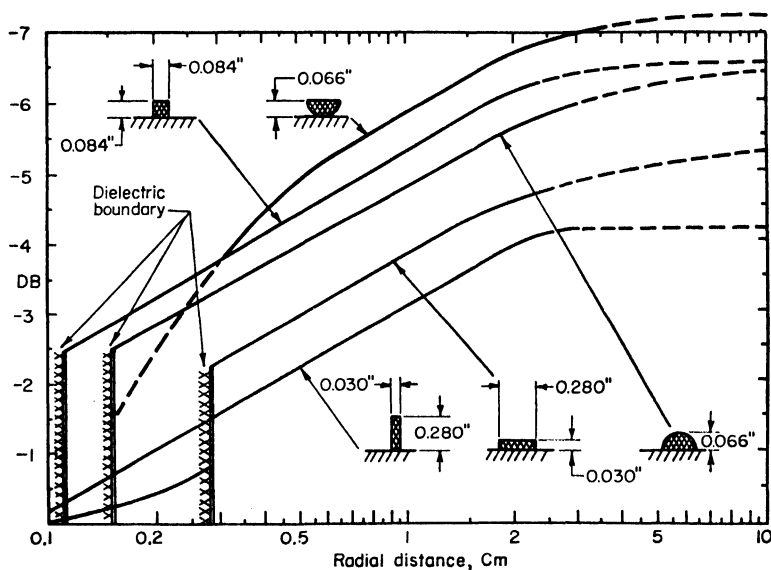



FIG. 78. Cross-sectional field distribution.

be operated in an image system. Figure 75 shows such a dielectric-image line in cross section and indicates the distribution of the E field. A metal ground plane of the width $2b$ carries a dielectric bead of semicircular cross section and $2a$ diameter.[1]

Components for dielectric-image-line systems have recently been developed. Several are shown by King[1] (see Figs. 76 to 78). While still in the early phases of development this system is expected to become very useful in the millimeter and submillimeter ranges.

[1] King, D. D., Properties of Dielectric Image Lines, *Trans. IRE*, **MTT-3** (2), 75, March, 1955. King, D. D., Circuit Components in Dielectric Image Lines, *Trans. IRE*, **MTT-3** (6), 35, December, 1955.

For right-angle bends, a reflector at exactly 45 deg and located at the intersection of the center lines of the straight sections, with sharply bent dielectric sections, yields much better match and lower losses than a gently curved line without the reflector. Figure 78 indicates cross-sectional field distributions for various cross sections of the dielectric bead. The maximum field occurs in the dielectric and is taken as the reference level. This figure indicates that the cross section of the dielectric bead is not very critical.

CAVITY RESONATORS

Conventional resonant circuits, consisting of a coil and capacitor, are hardly suitable at microwave frequencies, as the physical dimensions become too small and the losses by radiation may be considerable. At higher radio frequencies, sections of transmission line are used as circuit elements, and coaxial lines offer the particular advantage of minimizing radiation losses because the electromagnetic fields are completely enclosed within the outer conductor.

Coaxial line resonators are now generally included in the classification of cavity resonators, but the term did not come into general use until after it was realized that a hollow closed conducting box of arbitrary shape possessed electrical resonance properties similar to the conventional coil and capacitor circuit.[1] Rectangular, cylindrical, or spherical cavities may therefore be used at microwave frequencies to serve many of the same purposes that ordinary resonant circuits perform at lower frequencies, and in addition may possess an extraordinarily high efficiency. Qs in the order of many thousand are not uncommon. Moreover, all cavity resonators possess not one but an infinite number of resonant frequencies, analogous to the infinite number of transmission modes of a wave guide.

FIG. 79. Development of cavity resonator from a coil-capacitor circuit.

An illustration of how a cylindrical cavity may be developed from a conventional coil-and-capacitor circuit is given in Fig. 79. The coil reduces to a single turn of wire at higher frequencies; this wire is in turn paralleled by others to reduce the inductance and increase the resonant frequency, until finally a solid wall connects the capacitor plates, and the cavity resonator is formed.

With a cavity resonator, it is not possible in general to specify the exact inductance and capacitance of the equivalent resonant circuit. It is possible to specify, however, within limits of practical calculation the exact frequency of resonance for each of the resonance modes. Also the Q of a given cavity in a given mode may be uniquely defined. Finally, it is possible in many cases to specify the shunt impedance of a given cavity, once the path is determined along which the voltage shall be measured. In most instances, however, the choice of path is in some measure arbitrary.

38. Resonant Frequency. To calculate the resonant frequency of a cavity resonator, solutions to Maxwell's equations must be found which satisfy the boundary conditions imposed by the resonator. If the usual assumption is made that the cavity wall is a perfect conductor, the boundary conditions to be met are that no tangential electric field and no normal magnetic field exist at the surface of the cavity walls.

Under the above assumption, exact solutions are possible for resonators of simple geometrical shape. In practical cases, discrepancies between measured and calculated values for these simple shapes are extremely small, and usually attributable to mechanical imperfections in the cavity or to perturbations caused by the measuring device. A number of approximate methods of calculation have been developed which give more or less accurate answers for many more complex cavity shapes.

A principle of similitude may be applied to cavity resonators as follows: If all the linear dimensions of a cavity resonator are changed by a constant factor (and the resistivity of the walls changed by the same factor), the resonant wavelengths of all

[1] Hansen, W. W., A Type of Electrical Resonator, *J. Applied Phys.*, **9** (10), 654, October, 1938.

the normal modes in the cavity will be scaled by the same factor. The correction indicated in the parentheses is usually very small.

39. Resonator Q Values. The Q of a resonator in any given mode may be uniquely defined by the expression

$$Q = 2\pi \frac{\text{energy stored}}{\text{energy lost per cycle}} \tag{30}$$

If dielectric losses in the cavity resonator are neglected and if only losses resulting from conduction currents in the walls are considered, these losses are

$$\text{Energy lost per cycle} = \frac{\delta}{8} \int B^2 |d\sigma| \tag{31}$$

where δ = skin depth
B = magnetic field at wall of cavity
$d\sigma$ = element of area in cavity wall

The integral is carried out over the interior surface of the cavity. The energy stored in the cavity is

$$\text{Energy stored} = \frac{1}{8\pi} \int B^2 d\tau \tag{32}$$

with $d\tau$ an element of volume, and the integral carried out over the volume of the cavity. The Q of the cavity is then

$$Q = \frac{\lambda}{\delta} \frac{2}{\lambda} \frac{\int B^2 d\tau}{\int B^2 |d\sigma|} \tag{33}$$

It will be seen that Q is a dimensionless quantity. The above expression may be rewritten as

$$Q \frac{\delta}{\lambda} = \frac{2}{\lambda} \frac{\int B^2 d\tau}{\int B^2 |d\sigma|} \tag{34}$$

and the quantity $Q \frac{\delta}{\lambda}$ will be independent of frequency for a resonator of given shape and constant conductivity operating in a given mode. This quantity is sometimes known as the *form factor* of the cavity. The size of the resonator is inversely proportional to the frequency. As the skin depth varies as $\sqrt{\lambda}$, the Q of the cavity of given shape operating in a given mode will also vary as $\sqrt{\lambda}$, or inversely as $\sqrt{f}$.

For a nonreentrant resonator, the mean surface value of B^2 is approximately twice the mean value throughout the volume, and it can be said approximately that

$$Q \frac{\delta}{\lambda} \cong \frac{1}{\lambda} \frac{V}{S} \tag{35}$$

where V is the volume and S the bounding surface of the resonator, given in the same linear units as λ. For highly reentrant resonators, it is a better approximation to multiply the Q as given by Eq. (35) by a factor approaching 2. The form factor increases with increasing volume-to-surface ratio, and therefore large cavities which operate in one of the higher modes of oscillation will generally have higher Qs than smaller cavities operating in simpler modes at the same wavelength. Cavities which are highly reentrant have a low volume-to-surface ratio and are likely to have lower Qs than simpler shapes will give.

40. Shunt Impedance of Cavity Resonators. The shunt resistance, or shunt impedance at resonance, of a cavity resonator may be defined by

$$R_{sh} = \frac{\text{voltage}^2}{2 \times \text{energy lost per sec}} \tag{36}$$

where the voltage is the line integral of electric field along some reasonable path, usually the path along which the maximum value of voltage is developed. If the resonator is being driven by a beam of electrons, the path that is chosen in a calculation of shunt impedance is the path followed by the electron beam. The shunt impedance of a typical, nonreentrant cavity at a frequency of 3,000 Mc will be in excess of 1 megohm. If the cavity is reentrant, the shunt impedance will be lower, but values in excess of 100,000 ohms are typical for highly reentrant cavities such as are used in velocity modulation tubes at these frequencies.

If the wall conductivity of a given cavity is increased, both Q and the shunt resistance of the cavity will increase proportionally to the square root of the wall conductivity. Also, if a cavity of given shape is scaled in size to resonate at a different frequency, both Q and the shunt resistance will be inversely proportional to $\sqrt{f}$.

To say that a cavity has a high Q does not necessarily imply, however, that it also has a high shunt resistance, nor does a high shunt resistance imply a correspondingly high Q. It is possible to find cavity shapes in which a moderate Q is accompanied by an extremely high or an extremely low shunt resistance. A long cylindrical resonator operating in the $TM_{0,1,0}$ mode would have a high shunt resistance, and a reentrant coaxial resonator that is heavily capacitance-loaded would have a low shunt resistance.

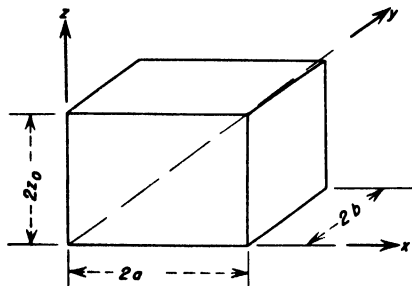



FIG. 80. Rectangular prism cavity resonator.

CHARACTERISTICS OF TYPICAL CAVITIES

41. Rectangular Resonators. The characteristics of a rectangular prism resonator, illustrated in Fig. 80, are readily calculated by analytical methods. A resonant wavelength will be found in such a resonator when

$$\lambda = \frac{4}{\sqrt{\left(\frac{l}{a}\right)^2 + \left(\frac{m}{b}\right)^2 + \left(\frac{n}{z_0}\right)^2}} \qquad (37)$$

where l = number of half-wave variations of field along the x axis
m = number of half-wave variations of field along the y axis
n = number of half-wave variations of field along the z axis
l, m, n = 0, 1, 2, 3, . . . but not more than one may equal zero for fields to exist

If the three dimensions of the resonator are equal and if the resonator is a perfect cube, there will be a twelvefold degeneracy; *i.e.*, 12 different modes will have identical resonant frequencies. If the three dimensions are unequal, this will be reduced to a twofold degeneracy, which in general will be destroyed by any irregularities in construction.

Counting each of the degenerate modes as a separate mode, the number of modes, dN, in a range $d\lambda$ is

$$dN = 8\pi \frac{V}{\lambda_1^4} d\lambda \qquad (38)$$

where V is the resonator volume and λ_1 is the center of the wavelength band $d\lambda$. The number of resonant states, N, in a large rectangular prism resonator with resonant wavelengths greater than some minimum wavelength λ_2 is given approximately by

$$N = \frac{8\pi}{3} \frac{V}{\lambda_2^3} \qquad (39)$$

This approximate formula is quite accurate even for low N.

For a resonator in which $a = b$ and where $l = m = 1$ and $n = 0$, the formulas for resonant wavelength, Q, and shunt impedance are given in Table 5, page 6-51. For large cubical resonators, operating in a high mode of oscillation, the Q is given approximately by

$$Q\frac{\delta}{\lambda} = \frac{a}{2\lambda_0} \tag{40}$$

42. Cylindrical Resonators.[1] For circular cylindrical resonators, as illustrated in Fig. 81, the infinite number of resonant modes that exist may be divided into two classes, corresponding to the two classes of transmission modes in a wave guide.

1. *TE Modes.* The electric field in these modes is everywhere transverse to the axis, while the magnetic field has an axial component. The resonant wavelength of these modes is given by

$$\lambda = \frac{4}{\sqrt{\left(\frac{l}{z_0}\right)^2 + \left(\frac{2u_{m,n}'}{\pi a}\right)^2}} \tag{41}$$

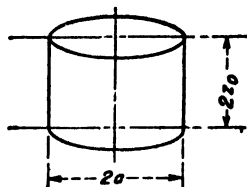



Fig. 81. Circular cylindrical cavity resonator.

Each of these modes occurs when the resonator is effectively a section of a circular wave guide that is an integral number of $\lambda/2$ long for some TE mode of transmission in the wave guide. The term l gives the number of $\lambda/2$ along the resonator axis and must therefore be an integral number. No modes exist in which $l = 0$. If the $TE_{m,n}$ mode in a wave guide is the mode being excited in the resonator, $u_{m,n}'$ is the nth root of the equation

$$J_m'(u') = 0 \tag{42}$$

Some of the lower roots of this equation are listed in Sec. 4.

The Q of the resonator for some of the lower modes may be obtained from Fig. 82.

The circular electric modes in circular cylindrical cavities, identified by subscripts of the form $TE_{0,n,l}$, are noteworthy for their exceptionally high Q. For this reason, they are ideal for high-precision cavity wavemeters, and also for echo boxes, as discussed in subsequent paragraphs.

2. *TM Modes.* The second class of resonant modes in circular cylindrical cavities is characterized by a magnetic field that is everywhere transverse to the axis, while the electric field has an axial component. The resonant wavelength of these modes is given by

$$\lambda = \frac{4}{\sqrt{\left(\frac{l}{z_0}\right)^2 + \left(\frac{2u_{m,n}}{\pi a}\right)^2}} \tag{43}$$

As before, each of these modes occurs when the resonator is effectively a section of circular wave guide that is an integral number of $\lambda/2$ long, this time for some TM mode of transmission in the wave guide. As before l gives the number of $\lambda/2$ along the axis and must therefore be an integral number, but in addition modes exist in which $l = 0$, corresponding to no transmission mode in a wave guide. These modes have an electric field that is everywhere parallel to the axis. If the $TM_{m,n}$ mode in a wave guide is the mode being excited in the resonator, $u_{m,n}$ is the nth root of the equation

$$J_m(u) = 0 \tag{44}$$

Some of the lower roots of this equation are listed in Sec. 4.

[1] Wilson, I. G., C. W. Schramm, J. P. Kinzer, High Q Resonant Cavities for Microwave Testing, *Bell System Tech. J.*, **25** (3), 408, July, 1946. See also Kinzer, J. P., and I. G. Wilson, End Plate and Side Wall Currents in Circular Cylindrical Cavity Resonator, *Bell System Tech. J.*, **26** (1), January, 1947.

The simplest mode in a cylindrical cavity is the $TM_{0,1,0}$ mode. For this mode the characteristics are listed in Table 5.

43. Spherical Resonators. The longest resonant wavelength will be found in a spherical cavity of radius a when

$$\lambda = 2.28a \tag{45}$$

and the next longest when

$$\lambda = 1.4a \tag{46}$$

The field configuration for these two modes is shown in Fig. 83. Values for Q and shunt impedance of the first mode are given in Table 5.

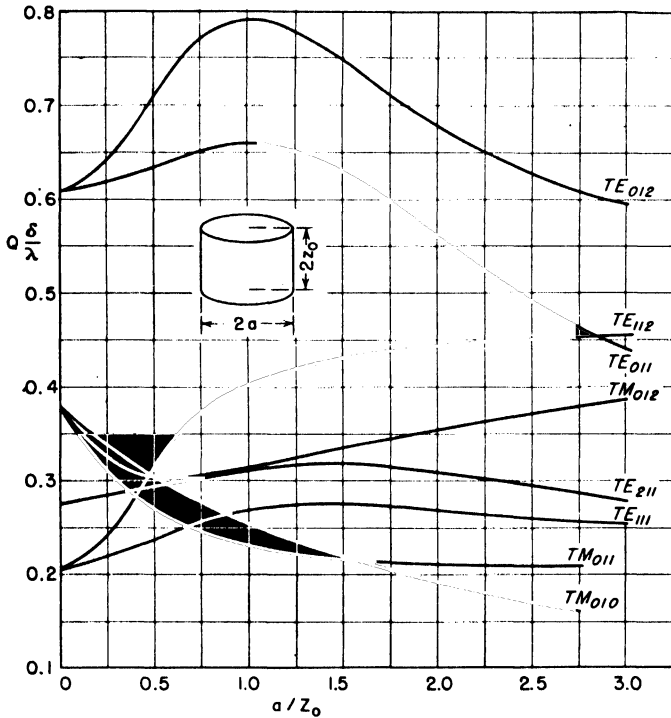
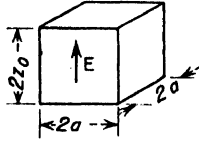
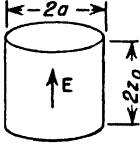
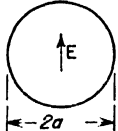



FIG. 82. Form factor of some of the lower modes in circular cylindrical resonators, from which Q may be determined.

44. Coaxial Line Resonators.[1] One type of coaxial line resonator is that shown in Fig. 84. The lowest *TEM* resonance in this cavity will be found when the cavity is $\lambda/2$ long. This resonant wavelength is given by $\lambda = 4z_0$. The Q of the cavity in this mode is

$$Q\frac{\delta}{\lambda} = \frac{1}{4 + \dfrac{2z_0}{b}\dfrac{1 + b/a}{\log_e b/a}} \tag{47}$$

[1] Barrow, W. L., and W. W. Mieher, Natural Oscillations of Electrical Cavity Resonators, *Proc. IRE*, **28** (4), 184, April, 1940. Some slightly misleading aspects of this paper are clarified by R. A. Kirkman and M. Kline, The Transverse Electric Modes in Coaxial Cavities, *Proc. IRE*, **34** (1), 33P, January, 1946.

Table 5. Characteristics of Simple Cavity Resonators Operating in the Lowest Mode of Oscillation

	Rectangular prism	Cylinder	Sphere
λ_0	2.828 a	2.61a	2.28 a
$Q\frac{\delta}{\lambda}$	$0.353 \dfrac{1}{1+\dfrac{a}{2z_0}}$	$0.383 \dfrac{1}{1+\dfrac{a}{2z_0}}$	0.318
$R\frac{\delta}{\lambda}$	$120\dfrac{z_0}{a} \dfrac{1}{1+\dfrac{a}{2z_0}}$	$144\dfrac{z_0}{a} \dfrac{1}{1+\dfrac{a}{2z_0}}$	104.4

and the highest Q is found when the diametric ratio is $b/a = 3.6$. The shunt impedance of this resonator is

$$R\frac{\delta}{\lambda} = \frac{60}{\pi}\frac{b}{z_0}\frac{\log_e^2 b/a}{1 + b/a}\,\frac{1}{1 + 2\dfrac{b}{z_0}\dfrac{\log_e b/a}{1 + b/a}} \tag{48}$$

The maximum shunt impedance is found when $b/a = 9.2$. Highest Q and highest shunt impedance require different values of diametric ratio, b/a. But neither shunt impedance nor Q is very critical to this ratio, as the Q will fall only to 78 per cent of its maximum value for $b/a = 9.2$, and the shunt impedance will fall only to 74 per cent of its maximum value for $b/a = 3.6$.

In addition to this lowest mode, resonant frequencies will be found whenever the cavity is an integral multiple number of $\lambda/2$ long, not only for the principal *TEM* mode, but also for any of the higher modes. None of these higher modes will give resonances unless the diameter of the line is sufficiently large. For the first of the higher modes

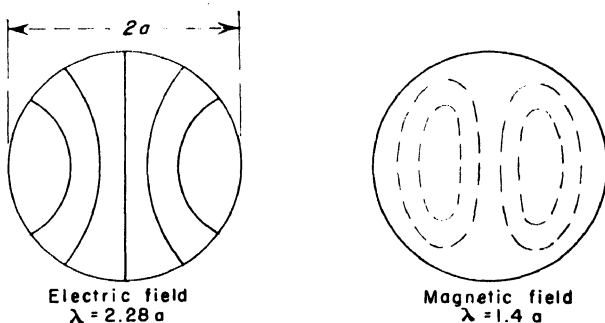



FIG. 83. Two lowest resonant modes in spherical cavity.

to be above cutoff and be a possible source of resonances, the arithmetic mean circumference of the line should exceed a full wavelength.

FIG. 84. Cavity resonator formed from a section of coaxial line shorted at both ends.

45. Capacitance-loaded Coaxial Line Resonators.[1] Resonators similar to the capacitance-loaded, coaxial line resonator shown in Fig. 85 are widely used in vacuum-tube u-h-f circuits. Klystron resonators are of this type in many instances, and disk-seal triodes are frequently employed with resonant circuits of this type.

Because of the relatively complex resonator geometry, accurate calculations of the resonant frequency and other characteristics are difficult to make, but approximate formulas are more readily developed.

Where the ratio of length to diameter of the cavity is large and the capacitance between the open end of the center conductor and the end of the cavity is small, the cavity will resonate when the inner conductor is approximately $\lambda/4$ long. As the capacitance at the open end of the center conductor becomes appreciable, the length of the center conductor for resonance becomes less. An approximate calculation of resonant wavelength which neglects the fringing fields at the capacitance gap gives for the resonant wavelength

$$\lambda = 2\pi \left(\frac{z_0 a^2}{2d} \log_e \frac{b}{a} \right)^{1/2} \tag{49}$$

The true resonant wavelength will always be greater than predicted by this formula, usually from 25 to 75 per cent greater. Accuracy will be best when the resonator is heavily capacitance-loaded.

For a reentrant resonator with considerable capacitance loading, the Q is given approximately by

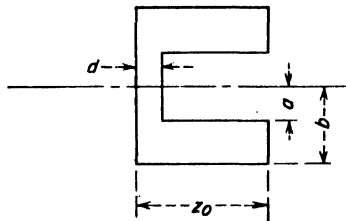
$$Q\frac{\delta}{\lambda} = \frac{2z_0}{\lambda} \frac{\log_e b/a}{2\log_e \dfrac{b}{a} + z_0\left(\dfrac{1}{b} + \dfrac{1}{a}\right)} \tag{50}$$

where z_0, b, and a are the cavity dimensions as given in Fig. 85. The shunt impedance of this cavity is approximately

$$R\frac{\delta}{\lambda} = 60\pi\left(\frac{2z_0}{\lambda}\right)^2 \frac{\log_e^2 b/a}{2\log_e \dfrac{b}{a} + z_0\left(\dfrac{1}{b} + \dfrac{1}{a}\right)} \tag{51}$$

FIG. 85. Capacitance-loaded, coaxial line resonator.

More exact calculations show that a resonator with fixed gap spacing d and inner radius a will have its maximum shunt impedance when the toroidal cross section is approximately square, *i.e.*, when $z_0 - d = b - a$.

If the cavity is relatively long and of small diameter, so that losses in the end plates may be neglected, but has appreciable loading capacitance, the Q is given approximately by

[1] Examples of approximate techniques of calculation are given in the following papers: Hansen, W. W., On the Resonant Frequency of Closed Concentric Lines, *J. Applied Phys.*, **10** (1), 38, January, 1939. Mayer, E., Resonant Frequencies of the Nosed-in Cavity, *J. Applied Phys.*, **17** (12), 1046, December, 1946. Hahn, W. C., New Method for the Calculation of Cavity Resonators, *J. Applied Phys.*, **12** (1), 62, January, 1941. Motz, H., Calculation of the Electromagnetic Field, Frequency, and Circuit Parameters of High-frequency Resonator Cavities, *J. IEE*, **39**, Pt. III (25), 335, September, 1946.

$$Q\frac{\delta}{\lambda} = \frac{\log_e \dfrac{b}{a}}{\pi\lambda\left(\dfrac{1}{b}+\dfrac{1}{a}\right)}\left(\frac{\sin\dfrac{4\pi z_0}{\lambda}}{\dfrac{z_0}{\lambda}+\dfrac{\sin\dfrac{4\pi z_0}{\lambda}}{4\pi}}\right) \tag{52}$$

and the shunt impedance approximately by

$$R\frac{\delta}{\lambda} = \frac{\log_e{}^2 b/a}{\lambda\left(\dfrac{1}{b}+\dfrac{1}{a}\right)}f(\lambda) \qquad \text{ohms} \tag{53}$$

where $f(\lambda)$ is given in Fig. 86.

Important progress toward a more rigorous analysis of cavity resonators is described in the recent literature.[1]

46. Coupling to Cavities. To utilize a cavity resonator, it is usually necessary to couple it to a transmission line. This may be done in a variety of ways. The coupling may be magnetic by means of a small coupling loop which links some of the magnetic flux lines of the desired mode in the resonator. The amount of coupling will be proportional to the magnetic flux enclosed by the loop and therefore to a first approximation proportional to the area of the loop.

To provide electric coupling to a cavity, the center conductor of a coaxial line may be extended as a probe into the cavity and will couple to the cavity if there is a component of electric field for the desired mode tangent to the probe. The coefficient of coupling will increase with increasing probe length and will also be greater if the probe is located at a point of high electric field strength.

FIG. 86. The function $f(\lambda)$ used in Eq. (53) in the computation of shunt impedance of a capacitance-loaded coaxial line resonator.

Cavities may be coupled to wave guides by holes through the metal wall that separates the guide from the cavity. There will be magnetic coupling through a round hole if the magnetic field at the surface of the guide has a component parallel to the magnetic field at the adjoining surface of the cavity. There will be electric coupling if both guide and cavity have components of electric field normal to the common surface between them. The coefficient of coupling in both instances will be proportional to the third power of the hole diameter, if the wall is thin and the hole is small.

If a narrow slot is used instead of a round hole for coupling, the electric coupling will be small, but there may be appreciable magnetic coupling if there is a component of magnetic field in both guide and cavity parallel to the slot. Coupling between cavities and transmission lines by these various means is illustrated in Fig. 87.

When a cavity is coupled to a transmission line, the resonant frequency and Q of the cavity are affected by the coupled load, and there is usually some question as to

[1] Whinnery, J. R., and D. C. Stinson, Radial Line Discontinuities, *Proc. IRE*, **43**, 46, January, 1955. Stinson, D. C., Resonant Frequencies of Higher-order Modes in Radial Resonators, *Trans. IRE*, **MTT-3**, 4, October, 1955. Schelkunoff, S. A., On Representation of Electro-magnetic Fields in Cavities in Terms of Natural Modes of Oscillation, *J. Applied Phys.*, **26**, 1231, October, 1955. Tinkham, M., and M. W. P. Standly, The Excitation of Circular Polarization in Microwave Cavities, *Proc. IRE*, **43**, 734, June, 1955.

just where the lines of demarcation are between the cavity, line, and coupling circuit. For many purposes it is convenient to regard the resonant frequency of the cavity under these conditions as that frequency at which a minimum standing-wave ratio will be measured on the transmission line, when energy is fed into the cavity. This is the frequency at which free oscillations will occur in the circuit if the transmission line is terminated in its characteristic impedance.

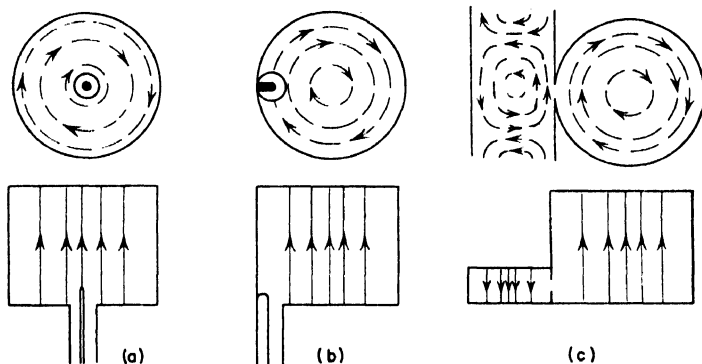



FIG. 87. Various methods of coupling between transmission lines and cavities. (*a*) Electric coupling between coaxial line and cavity by probe. (*b*) Magnetic coupling between coaxial line and cavity by loop. (*c*) Magnetic coupling between rectangular wave guide and cavity by a hole in the common wall.

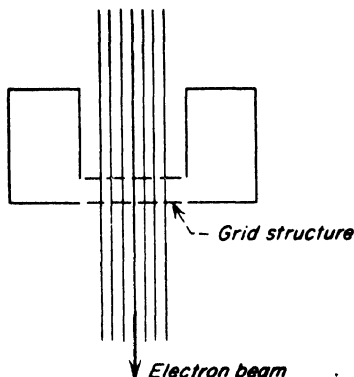
The amount which the cavity is loaded by the coupled transmission line may be determined by the standing-wave ratio at resonance in the line feeding the cavity. The following Qs have been defined which are applicable to this situation:

$$\text{Unloaded } Q = Q_0 = \frac{2\pi f \times \text{energy stored in cavity}}{\text{power dissipated in cavity}}$$

$$\text{External } Q = Q_e = \frac{2\pi f \times \text{energy stored in cavity}}{\text{power dissipated in load}}$$

$$\text{Loaded } Q = Q_L = \frac{2\pi f \times \text{energy stored in cavity}}{\text{power dissipated in load and cavity}}$$

The various Qs are then related by the following equations:

$$\frac{Q_0}{Q_L} = 1 + \beta \qquad (54)$$

$$\frac{Q_0}{Q_e} = \beta \qquad (55)$$

$$\frac{Q_e}{Q_L} = 1 + \frac{1}{\beta} \qquad (56)$$

FIG. 88. An electron beam coupled to a cavity.

The term β is equal to the voltage standing-wave ratio at resonance in the input transmission line, if the positions of the minimums of the standing waves at resonance coincide with the positions of the maximums when the cavity is detuned. If the standing-wave minimums at resonance coincide with the minimums when the cavity is detuned, β is the reciprocal of the standing-wave ratio measured at resonance on the input line.

Energy may be supplied to a cavity by an electron beam passing through the cavity. If the transit time through the resonator is short compared to a cycle, there will be power transferred from the beam to the cavity if the beam has an alternating component of current, *i.e.*, if the beam is in the form of bunches of electrons. These bunches

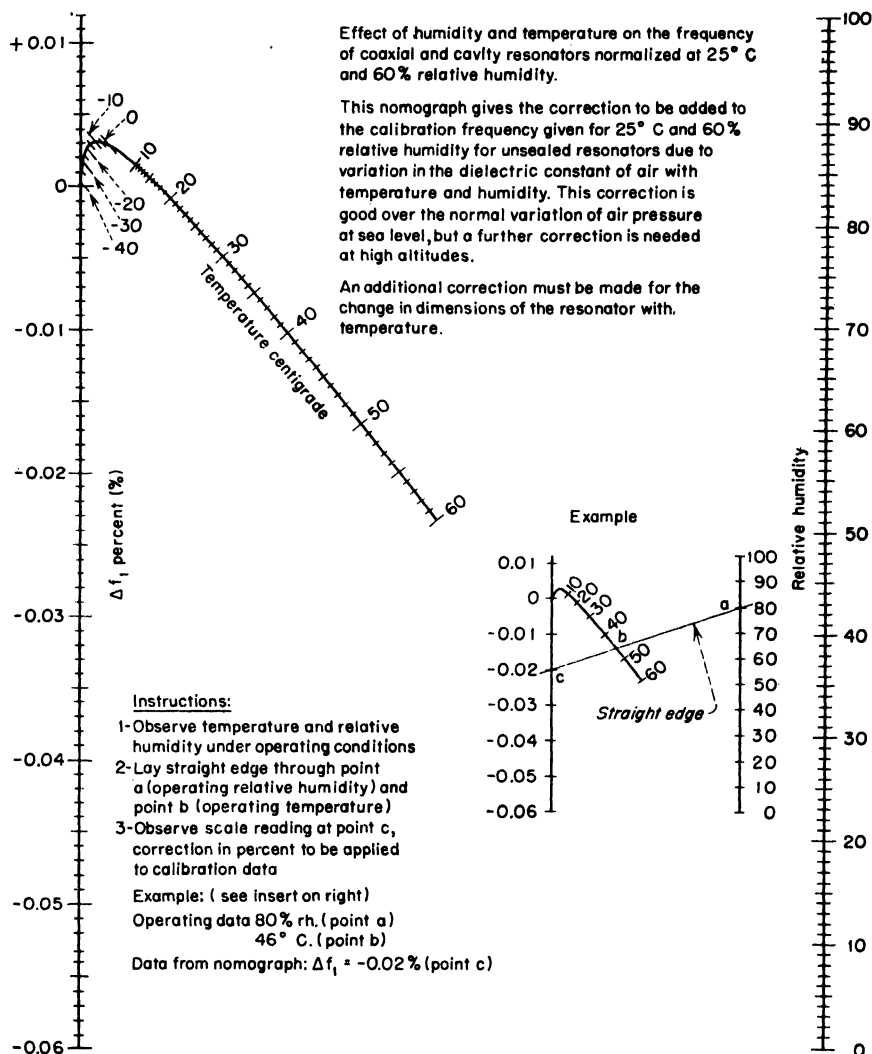



FIG. 89. Nomogram for calculating effect of humidity on resonant-line or cavity wavemeters.

may be formed either by velocity modulation action coupled with passage through a drift space, as in a klystron, or by the action of an alternating voltage applied to a control grid, as in a class C amplifier. The alternating voltage developed across the path of the beam will be equal to the product of resonator shunt impedance and the alternating component of beam current, provided the peak resonator voltage does not

exceed the beam voltage. A highly reentrant cavity, as shown in Fig. 88, is usually required to give an appreciable shunt impedance coupled with short transit time.

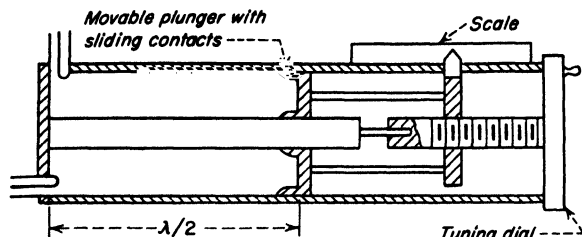



Fig. 90. Wavemeter cavity formed from section of coaxial line, shorted at both ends, of variable length.

If the transit time through the cavity is appreciable compared to a cycle, calculation of the power transfer is more difficult. It can be shown that a beam which is unmodulated when entering a cavity is capable of exciting oscillations in the cavity if the transit time through the cavity is of the proper value.[1]

47. Effect of Temperature and Humidity upon Resonant Frequency. The resonant frequencies of a cavity depend upon the physical dimensions of the cavity, and if these dimensions vary with changing temperature, the resonant frequencies will also change. If the cavity is constructed of a single metal, the percentage change in resonant wavelength for any mode will be equal to the percentage change in linear dimensions, which in turn will depend upon the temperature change and the expansion coefficient of the metal.

Fig. 91. Coaxial cavity wavemeter which is resonant when the center conductor is approximately λ/4 long.

The resonant frequencies of the cavity will also vary if the dielectric constant of the medium inside the cavity changes. If the cavity is not sealed, the changing dielectric constant of the air resulting from variations in temperature and humidity will cause changes in its resonant frequencies. These changes may be determined from the nomogram of Fig. 89.[2]

Fig. 92. Cavity wavemeter in which cavity is cross between λ/4 coaxial line resonator and a circular cylindrical resonator operating in the $TM_{0,0}$ mode.

48. Cavity Wavemeters. The extremely high Qs associated with cavity resonators make them ideal for use as wavemeters at microwave frequencies. The simplest type of cavity wavemeter, which is really a coaxial line Lecher wire system, is shown in Fig. 90. The cavity is a section of coaxial line, shorted at both ends, which is resonant when λ/2 long. The resonant frequency is varied by moving the shorting plunger at one end of the cavity to change the length of the cavity.

A second type of coaxial cavity that is widely used is shown in Fig. 91. This cavity is a λ/4 coaxial line, open at one end and shorted at the other. The resonant frequency is tuned by changing the length of the

[1] Marcum, J., Exchange of Energy between an Electron Beam and an Oscillating Electric Field. *J. Applied Phys.*, **17** (1), 4, January, 1946.
[2] Gaffney, F., Microwave Measurements and Test Equipments, *Proc. IRE*, **34** (10), October, 1946.

plunger forming the center conductor. A variation of this design is shown in Fig. 92.[1] This cavity is really a hybrid between a $\lambda/4$ coaxial line resonator and a cylindrical resonator operating in the $TM_{0,1,0}$ mode, resembling the former when the plunger is extended, and the latter when the plunger is nearly withdrawn. This design offers advantages of higher Q coupled with a tuning characteristic that is nearly linear in frequency.

The $TE_{1,0}$ mode in circular wave guide has an exceptionally low attenuation, and wavemeters employing this mode have a correspondingly high Q. A typical wavemeter design employing this mode is shown in Fig. 93. The resonant frequency is changed by moving the shorting plate at one end of the cavity. This plate is not required to make contact with the cavity walls, as with this mode of operation there is no current flowing across this gap from the side walls to the end plates. A cavity operating in this mode is capable of giving spurious indications by resonating in modes other than the desired $TE_{0,1,1}$ mode, and these spurious indications must be eliminated by damping out the undesired modes. This may be done with the slug of lossy dielectric shown mounted behind the tunable end plate in Fig. 93. For the undesired modes, currents will flow from the side walls to the end plate, and fields will penetrate into the space behind the tunable end plate. Energy is absorbed from these fields by the lossy dielectric, and the Q of the undesired modes is correspondingly lowered. Fields from the desired mode do not penetrate into this space, and the Q is therefore unaffected by the lossy dielectric.

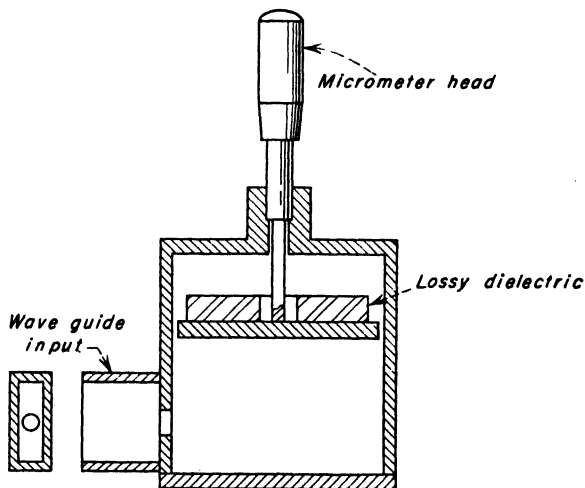



FIG. 93. High-Q cylindrical-cavity wavemeter operating in $TE_{0,1,1}$ mode. The lossy dielectric behind the tunable end plate eliminates spurious resonances.

49. Echo Boxes.[2] An echo box is a piece of test equipment used for over-all checks of radar systems. It consists of a high-Q cavity to which is loosely coupled a pickup antenna and also a crystal detector (Fig. 94). The echo box is placed in front of the antenna of the radar system, so that part of the energy radiated from the radar antenna during the transmitter pulse is picked up by the echo-box antenna and stored in the cavity in the form of oscillating electromagnetic fields. When the transmitter pulse is completed, this energy stored in the cavity will be reradiated through the echo-box antenna and detected by the radar receiver. The strength of this reradiated signal will decrease exponentially with time, and the length of time during which it can be detected by the radar receiver is a measure of the over-all performance of the radar system. This time, called the *ring time* of the echo box, depends upon the system tested.

For the rate of signal decay to be minimized, the Q of the echo-box cavity should be as high as possible. Two general types of cavities, tuned and untuned, are in use. The untuned cavity has dimensions large compared to a wavelength, and the various

[1] Essen, L., Cavity-resonator Wavemeters, *Wireless Engr.*, **23** (272), 126, May, 1946.

[2] Gaffney, *loc. cit.* Also Green, E. E., H. J. Fisher, and J. G. Ferguson, Techniques and Facilities for Microwave Radar Testing, *Bell System Tech. J.*, **25** (3), 435.

resonant modes are therefore spaced so closely in frequency that one or more will be excited by the frequency spectrum of the radar transmitter. For this reason it is unnecessary to tune the cavity to the frequency of the transmitter. A typical untuned echo box for a frequency of 3,000 Mc will be a nearly cubical box, approximately a meter on each side. A perfect cube is avoided to prevent mode degeneracy, as discussed in Sec. 41.

A tuned echo box is smaller and for that reason more convenient to use, but it must be tuned to the frequency of the radar system. To obtain the requisite high Q

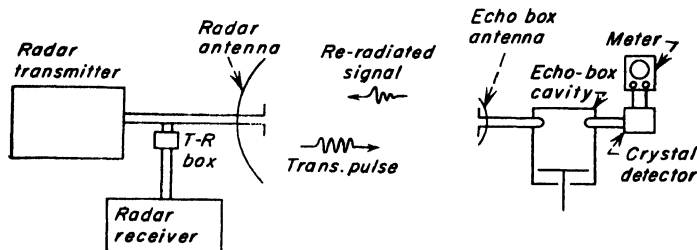



FIG. 94. Echo box being used for a check of the over-all performance of a radar system.

in a small cavity, the operating mode must be carefully chosen, and circular electric modes in circular wave guides are generally preferred. At 3,000 Mc, a cavity operating in the $TE_{0,1,1}$ mode has been found satisfactory with an operating Q over 40,000. At higher frequencies the ring time associated with a cavity of this design decreases rapidly, and it is therefore necessary to go to higher modes to get satisfactory performance. Much work has been done with cavities operating in the $TE_{0,1}$ mode in circular wave guide, but between 10 and 20 $\lambda/2$ long. These cavities give Qs well in excess of 100,000, but many troubles of interaction between the desired and undesired modes are encountered, which require very careful design to overcome.

References

Brainerd, J. G., *et al.*: "Ultra-high-frequency Techniques," Van Nostrand, 1942.

King, R. W. P., H. R. Mimno, and A. H. Wing: "Transmission Lines, Antennas, and Wave Guides," McGraw-Hill, 1945.

Marcuvitz, N.: "Waveguide Handbook," M.I.T. Radiation Laboratory Series, 10, McGraw-Hill, 1951.

Montgomery, C. G.: "Technique of Microwave Measurements," M.I.T. Radiation Laboratory Series, 11, McGraw-Hill, 1948.

——— *et al.*: "Principles of Microwave Circuits," M.I.T. Radiation Laboratory Series, 8, McGraw-Hill, 1948.

Moreno, T.: "Microwave Transmission Design Data," McGraw-Hill, 1948.

Pound, R. V., and E. Durand: "Microwave Mixers," M.I.T. Radiation Laboratory Series, 16, McGraw-Hill, 1948.

Ragan, G. L.: "Microwave Transmission Circuits," M.I.T. Radiation Laboratory Series, 9, McGraw-Hill, 1948.

Ramo, S., and J. R. Whinnery: "Fields and Waves in Modern Radio," Wiley, 1944.

Reich, Herbert J.: "Very High Frequency Techniques," Vols. I and II, McGraw-Hill, 1947.

Sarbacher, R. I., and W. A. Edson: "Hyper and Ultra High Frequency Engineering," Wiley, 1943.

Schelkunoff, S. A.: "Electromagnetic Waves," Van Nostrand, 1943.

Skilling, H. H.: "Fundamentals of Electric Waves," Wiley, 1942.

Slater, J. C.: "Microwave Transmission," McGraw-Hill, 1942.

Smullin, L. D., and C. G. Montgomery: "Microwave Duplexers," M.I.T. Radiation Laboratory Series, 14, McGraw-Hill, 1948.

Staff of the Radar School at M.I.T.: "Principles of Radar," McGraw-Hill, 1946.

Van Voorhis, S. N.: "Microwave Receivers," M.I.T. Radiation Laboratory Series, 23, McGraw-Hill, 1948.

Young, V. J.: "Understanding Microwaves," Rider, 1946.

CHAPTER 7

ELECTRON TUBES FOR LOW AND MEDIUM FREQUENCIES

By George D. O'Neill[1]

The present chapter reviews the basic principles on which low- and medium-frequency tubes are designed and made.

There are several classes of devices which, on this basis alone, would require inclusion if further restrictions were not imposed. Among these would be certain cold-cathode, gas-filled tubes, thyratrons, pool-type rectifiers, and television pickup tubes. Although some of these devices are of considerable interest, discussion of their characteristics and design is omitted largely on the ground that information in greater detail than is given in the manufacturers' ratings is of value to a much smaller group of readers than is the case for high-vacuum electron tubes.

Except in a few instances, methods of testing for the various characteristics of electron tubes are not described, primarily for the reason that the standards of the IRE[2] are widely available and more complete than could be given here.

The extent to which various topics in this chapter are pursued depends largely upon relative need. By way of example, thermionic emission is the concern of more engineers than is photoelectric emission; more engineers are concerned with the processing, design, manufacture, and use of receiving tubes than of transmitting tubes.

Finally, the manner of treatment of the various topics is, it is hoped, such as to be of maximum benefit to the average engineer, visualized as having a fair background in physics and mathematics but without the advantages of graduate training or the equivalent in experience. Discussion of vacuum techniques and certain "practical" aspects of tube making such as glass sealing, exhausting, and coating techniques have not been included for the reason that no amount of reading on such topics could take the place of learning by even a small amount of experience in a laboratory or factory.

ELECTRON EMISSION

The passage of electrons from a solid into a gas or vacuum is called *electron emission*. The mechanism by which emission occurs in most cases classifies the general phenomenon into the categories thermionic, photoelectric, secondary, and field emission.

1. Thermionic Emission from Metals. The kinetic energies of the free electrons in a metal are distributed in accordance with the Fermi-Dirac relation shown in Fig. 1. Here the abscissas represent kinetic energy on a linear scale with zero energy at the origin. The ordinates may be taken as the probability of finding an electron with the corresponding energy.

Curve *a* shows the energy distribution of the free electrons at absolute zero. It is noted that the most probable energy corresponds to the voltage E_1 and that there are no electrons having energies higher than this amount. At a higher temperature—say, 293°K or room temperature—the curve will look more like *b*; it is seen that a few electrons now have energies greater than E_1. Curve *c* represents the energy distribution at a still higher temperature.

The energy E_2 of Fig. 1 represents a critical level. An electron arriving at the surface of the metal may emerge from it if the surface-directed component of its energy

[1] Research Laboratories, Sylvania Electric Products Inc.

[2] Standards on Electron Tubes: Methods of Testing, *Proc. IRE*, **38**, Part I (8), 917, August, 1950; **38**, Part II (9), 1079, September, 1950.

is equal to or greater than E_2. If the surface-directed component is just equal to E_2, the electron emerges with zero initial velocity; if it is greater than E_2 the excess is the measure of the initial velocity upon emergence.

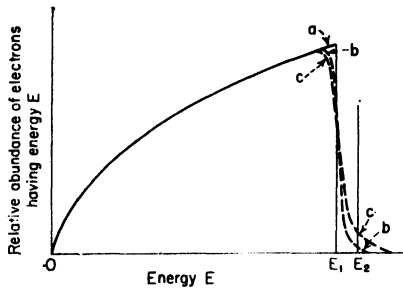



Fig. 1. Energy distribution of electrons in metals.

Since an electron may have maximum energy measured by E_1 volts at absolute zero, 0°K, the difference between E_2 and E_1 measures the amount of energy that must be added by heating. This energy difference $(E_2 - E_1)$ is equal to $e\phi$, where e is the electron charge and ϕ is called the *work function.*

Not all electrons having a surface-directed energy greater than the work function are emitted; a fraction $\bar{r}$ will be reflected from the surface.

The quantitative relation between these variables is expressed by the relation

$$I = A_0(1 - \bar{r})T^2\epsilon^{-b/T} \qquad (1)$$

where I = emission current per unit area

A_0 = constant having a theoretical value of 120 amp/sq cm per deg² for metals; experimental values are usually about half this figure

$\bar{r}$ = reflection coefficient

T = absolute temperature

The exponent b in Eq. (1) is a measure of the work required to remove an electron from the cathode and is proportional to the work function of the cathode, the relation being

$$b = \phi e/k \qquad (2)$$

where ϕ = work function, volts

e = magnitude of charge of electron

k = Boltzmann's constant

Expressed in practical units,

$$b = 1.16 \times 10^4\phi \quad \text{deg K} \qquad (3)$$

For most practical cases, therefore, the thermionic-emission density for metals may be taken as

$$I = 60T^2\epsilon^{\ 11,600\phi/T} \quad \text{amp/sq cm} \qquad (4)$$

when T is given in Kelvin or absolute degrees centigrade and ϕ is in volts.

Agreement between theory and practical results is usually satisfactory for metals, since a check on the value of ϕ from thermionic-emission data can be made by determining the long-wavelength limit of the photoelectric effect for the same cathode. With composite surfaces a number of factors commonly lead to inconsistencies, and recourse is generally had to statistical data where the ultimate objective is the design of an electron tube.

2. Tungsten Filaments. A brief summary of some practical thermionic cathode data is given below. These data are not, of course, intended to represent more than a small part of the information available in the complete literature.

Tungsten. Pure tungsten filaments were extensively used in early electron tubes. The ratio of emission to heating power is very low on account of the high work function of tungsten; consequently now only large transmitter tubes, where conditions of temperature and voltage are severe, employ tungsten.

The work function ϕ is given as 4.54 volts. Accordingly, the emission I_s in amperes per square centimeter may be calculated by Eq. (4) with the results tabulated for deg K, shown in the fourth column of Table 1.

Table 1. Properties of Tungsten and Thoriated Tungsten Wire

Temp, °K	ρ, ohm-cm × 10^6	η, watts/cm²	I_s (tungsten), amp/cm²	I_s (thoriated tungsten), amp/cm²	Expansion ($L_T/L_{293°K}$)	I' ($I/d^{3/2}$)	$E' \times 10^3$ ($Ed^{1/2}/L$)
293	5.49	....	....	....	1.0000	0	0
500	10.56	9.71×10^{-3}	....	....	1.0009	47.62	0.640
1000	24.93	0.602	....	9.1×10^{-8}	1.0032	244.1	7.749
1100	27.94	1.03	....	2.5×10^{-6}	1.0037	301.0	10.71
1200	30.98	1.66	....	4.0×10^{-5}	1.0042	363.4	14.34
1300	34.08	2.57	....	4.0×10^{-4}	1.0048	430.9	18.70
1400	37.19	3.83	....	3.1×10^{-3}	1.0053	503.5	23.85
1500	40.36	5.52	9.4×10^{-8}	1.8×10^{-2}	1.0059	580.6	29.85
1600	43.55	7.74	9.27×10^{-7}	8.5×10^{-2}	1.0065	662.2	36.73
1700	46.78	10.62	7.08×10^{-6}	3.4×10^{-1}	1.0071	747.3	44.52
1800	50.05	14.19	4.47×10^{-5}	1.15	1.0078	836.0	53.28
1900	53.35	18.64	2.28×10^{-4}	3.55	1.0084	927.4	63.02
2000	56.67	24.04	1.00×10^{-3}	9.55	1.0090	1,022	73.75
2100	60.06	30.5	9.93×10^{-3}	23.44	1.0097	1,119	85.57
2500	73.91	69.8	0.298	....	1.0125	1,526	143.6
3000	95.76	185.8	14.15	....	1.0164	2,072	243.0
3655*	117.1	382.6	479.9	....	1.0221	2,838	423.4

* Melting point.

Thoriated Tungsten. The inclusion of a few parts per hundred of thorium oxide in tungsten wire permits, with proper processing, a considerable reduction in the work function. This processing consists in conversion of some of the thorium oxide to thorium metal which then diffuses outward, covering the tungsten. The emission from such a filament is given by Eq. (1), in which

$$b = 52{,}600 - 18{,}800\theta' \quad \text{deg} \tag{5}$$

where the factor θ' is related to θ, the fraction of the surface covered with a monatomic layer of thorium, as in Table 2; $A_0(1 - \bar{r})$ is taken as 60.2 amp/cm²/deg².

Table 2

θ	θ'	θ	θ'
0.00	0.00	0.5	0.790
0.05	0.127	0.6	0.863
0.1	0.240	0.7	0.920
0.2	0.430	0.8	0.960
0.3	0.579	0.9	0.988
0.4	0.697	1.0	1.000

It will be recognized that the temperature equivalent 52,600 in Eq. (5) is the value of b for pure tungsten. Therefore, for no thorium the work function is that of tungsten, 4.54 volts; for the fully activated surface, the work function will be about 2.9 volts.

3. Activation of Thoriated Tungsten Cathodes. Activation—the process of causing the thorium to diffuse out of the wire and coat the surface—is usually accomplished in one of two ways:

For *small filaments*, in which the operating current is of the order 0.25 amp or less, thorium oxide is reduced by raising the temperature of the filament to about 3000°K by the application of a voltage corresponding to two to three times that required for

normal operation. After about 30 sec the voltage is reduced to about 1.5 times the normal value and held for about 10 to 15 min to permit diffusion of thorium to the surface of the wire. The word "about" has been used advisedly, since some experimentation is usually in order before a satisfactory "aging" schedule for a particular design of filament can be specified.

For *large filaments*, as in transmitting tubes, the thorium oxide is not sufficiently reduced by the heat-treatment described above; consequently resort is had to carburizing, in which the filament is glowed at high temperature in the presence of a hydrocarbon gas such as acetylene at very low pressure (or in a heavier compound diluted with hydrogen at atmospheric pressure). It is usually preferred that the pressure of the gas be restricted to an amount just sufficient to cause penetration of the carbon to the desired depth; a safer practice is to admit a still lower quantity and remove the resultant hydrogen before more acetylene is admitted. Carburization in the presence of an excessive amount of gas frequently results in uneven penetration and excessive brittleness. The acetylene must be highly purified and entirely free of oxygen or water vapor. Ordinarily, the carburization is carried to the depth representing about 20 per cent of the cross-section area. In this process the outer shell of the filament is converted to a very brittle carbide of tungsten having high electrical resistivity. During carburization the thorium oxide is reduced to metallic thorium which may be diffused to the surface by an aging schedule similar to that described for the small filaments.

4. Properties of Pure and Thoriated Tungsten Cathodes. Column 5 of Table 1 gives the emission of pure and thoriated tungsten. Pure tungsten filaments are ordinarily operated at temperatures of 2500 to 2570°K, thoriated tungsten at 1900 to 2000°K.

Table 1 provides data for the design of a tungsten filament operating in vacuum. Excluding second-order effects, such as temperature of the surroundings and the variation in temperature along the filament, the variables will be

L = length of filament
E = filament voltage
I = filament current
ρ = resistivity
η = radiation intensity
r = radius of wire
d = diameter of wire

Formulas have been published for corrections due to cooling of the filament by the lead wires and hooks. Since the complete data necessary for precise determination of these corrections are seldom available, it is generally preferable, for ordinary computations, to follow a rule of thumb. This is simply to observe, in a sample or similar tube, the amount of filament which operates below visible incandescence and to consider that this part has zero resistance while the remainder has uniform resistance per unit length. The hot part of the filament is computed and the parts cooled by leads and hooks then added to give the final length. In the paragraphs to follow, L means the length of the hot part of the filament.

The hot resistance of the filament will be taken as the quotient of total filament voltage E by filament current I, and is related to the resistivity and dimensions by

$$E/I = \rho L/\pi r^2 \quad \text{ohms} \tag{6}$$

With the filament heated by power EI, the radiation intensity η, or power radiated per unit area, is

$$\eta = EI/2\pi rL \tag{7}$$

Since Table 1 gives ρ and η for centimeter units, L and r will be expressed in centimeters and the η will be in watts per square centimeter.

From Eqs. (6) and (7), the current and voltage of a wire 1 cm long and 1 cm in diameter may be computed for the temperatures given in the table, using the corresponding values of ρ and η. These values, I' and E', are given in the last two columns of the table. For values of I' and E' at temperatures intermediate between those

given, interpolation between the next higher and lower values yields results which are sufficiently accurate for most purposes. Then, for a wire of any other diameter d and length L operating at the same temperature, the current I will be

$$I = I'd^{3/2} \quad \text{amp} \tag{8}$$

and E will be

$$E = E'L/d^{1/2} \quad \text{volts} \tag{9}$$

In practice, the temperature of a wire having known dimensions is determined by computation of I' from current and diameter or of E' from length, voltage, and diameter. The temperature is then found from the table. In design, the length, voltage, and temperature are likely to be specified, in which case the diameter will be obtained from Eq. (9) by substituting the value of E' corresponding to the desired temperature. This diameter is then substituted into Eq. (8) with the corresponding value of I' to obtain the required current.

The constants of Table 1 apply only to clean tungsten or thoriated tungsten; hence the values of η, E', and I' do not apply to the heater wire of an indirectly heated cathode. However, Eq. (6) does apply in any case and is of use in computing the temperature of heater wire, when the design is achieved through cut-and-try methods. Here ρ is determined from voltage, current, and dimensions of the experimental tubes. The final tube should be so designed that the operating temperature of the heater does not exceed about 1450°K. Temperatures in excess of this value are apt to result in burnout trouble during exhaust or aging and may result in early failure of the heater.

It is difficult to measure with the degree of accuracy required the average diameter of a small tungsten wire when the wire is to be used in an electron tube or incandescent lamp, and recourse is had to the weight as measured by a sensitive torsion balance. Except for large sizes, the wire is commonly purchased and specified in terms of its weight in milligrams for a 200-mm length, the tolerance being specified—commonly 2 per cent. Table 3 gives the data required for the conversion of diameter in millimeters and mils (thousandths of an inch) to weight in milligrams per 200 mm.

5. Carburized Filament Design. In designing a carburized filament, it may be considered that the electron emissivity is the same as for the uncarburized thoriated wire. The resistance depends on the depth of carburization. As an approximation, the carburized portion may be taken to have infinite resistivity; *i.e.*, in the computations the resistance is taken on the basis of the uncarburized core while the radiation is taken for the original diameter. At the normal operating temperature of 1900 to 2000°K, the radiation intensity of the carburized wire is about 20 per cent greater than for clean wire at the same temperature. Thus, when the wire is 20 per cent carburized, a given current heats both to approximately the same temperature, the voltage being about 20 per cent greater for the carburized filament.

6. Oxide-coated Cathodes. The oxide-coated or Wehnelt cathode comprises a base metal such as nickel, tungsten, or one of a large number of alloys, coated with the oxides of barium and strontium or of barium, strontium, and calcium. There are numerous other types of coated cathodes, but their commercial importance is small relative to that of the BaO-SrO variety at the time this is written.

The electronic nature of an oxide is quite different from that of a metal; consequently the rather simple model of an emitter discussed in Sec. 1 must be considerably altered if a useful picture of the mechanism of electron emission from a coated cathode is to be obtained.

In Eq. (1) the quantity A_0 is derived on the basis that there are as many "free" electrons in the metal as there are atoms. The oxide coating of a thermionic cathode is an n-type semiconductor, in which conduction and emission are made possible by electrons "donated" by the ionization of impurities. These impurities may be chemical in nature, such as barium atoms not occurring as a normal point in the crystal lattice, or they may be defects such as vacancies in the latter. Consequently the number of conduction electrons in a crystal or grain of coating is far smaller than

the number of atoms. The result is that a term such as A_0, if used in connection with an oxide-coated emitter, is not a constant but depends on the composition, condition, and temperature of the coating.

Furthermore, the meaning of the term b, or $\phi e/k$, as used in Eq. (1), is not quite the same as one obtains from Fig. 1. An explanation of the concept of work function as applied to a semiconductor is beyond the scope of this book, but it should be pointed out that the net effect, from a practical point of view, is that the thermionic current obtainable from a coated cathode varies inversely with a power of the work function about as indicated by the equation. In a properly prepared, well-aged coated cathode, the work function $e\phi$ has a value of about 1.0 to 1.05 electron volts as compared with 4.54 for tungsten and 2.9 for thoriated tungsten; it is this fact which explains why the emission is so much higher at the same temperature.

In the case of receiving tubes employing indirectly heated cathodes, the power supplied by the heater is usually about 2.75 to 3.0 watts per sq cm of cathode area; in high-vacuum rectifiers it is higher, sometimes as great as 4.0 watts per sq cm to provide for the high peak currents frequently required.

Table 3. Weight and Diameter of Tungsten Wire

Weight, mg/200 mm	Diameter, mm	Diameter, mils	Weight, mg/200 mm	Diameter, mm	Diameter, mils
1	0.01830	0.72	32	0.1035	4.08
1.5	0.02242	0.88	33	0.1050	4.14
2	0.02589	1.02	34	0.1066	4.20
2.5	0.02894	1.14	35	0.1083	4.26
3	0.03167	1.25	36	0.1097	4.32
3.5	0.03425	1.35	37	0.1113	4.38
4	0.0366	1.44	38	0.1127	4.44
4.5	0.0388	1.53	39	0.1142	4.50
5	0.0409	1.61	40	0.1157	4.56
5.5	0.0429	1.69	41	0.1172	4.61
6	0.0448	1.765	42	0.1185	4.67
6.5	0.0467	1.84	43	0.1199	4.72
7	0.0484	1.905	44	0.1213	4.76
7.5	0.0501	1.97	45	0.1227	4.83
8	0.0517	2.04	46	0.1240	4.89
8.5	0.0533	2.10	47	0.1254	4.94
9	0.0549	2.16	48	0.1267	4.99
9.5	0.0564	2.22	49	0.1280	5.04
10	0.0578	2.28	50	0.1294	5.09
11	0.0607	2.39	55	0.1357	5.34
12	0.0634	2.495	60	0.1417	5.58
13	0.0659	2.60	65	0.1474	5.81
14	0.0684	2.69	70	0.1531	6.03
15	0.0708	2.79	75	0.1584	6.23
16	0.0732	2.88	80	0.1536	6.44
17	0.0754	2.97	85	0.1687	6.64
18	0.0776	3.055	90	0.1735	6.83
19	0.0798	3.14	95	0.1783	7.02
20	0.0818	3.22	100	0.1830	7.20
21	0.0838	3.30	110	0.1919	7.56
22	0.0858	3.38	120	0.2005	7.89
23	0.0878	3.455	130	0.2087	8.22
24	0.0897	3.53	140	0.2165	8.53
25	0.0915	3.60	150	0.2242	8.83
26	0.0935	3.67	160	0.2314	9.11
27	0.0951	3.75	170	0.2387	9.39
28	0.0968	3.81	180	0.2455	9.67
29	0.0985	3.88	190	0.2523	9.93
30	0.1002	3.95	200	0.2589	10.18
31	0.1018	4.01			

7. Emission Efficiency. It is difficult to decide what one would call an "average" oxide-coated cathode. On the basis of usual experience, a fair comparison might be made of the emission efficiencies—the quotient of emission current in amperes by heating power in watts—between the three principal types of thermionic cathodes. The quantities given are taken at the ordinary level of emission density of an oxide-coated receiving tube cathode.

Emitter	I_s/sq cm	Watts/sq cm	I_s/watt
Oxide-coated	0.10	2.75	0.0364
Thoriated tungsten	0.10	8.0	0.0125
Tungsten	0.10	55	0.0018

When the emission is measured by applying short pulses of voltage, for example, 1-μsec pulses at a repetition rate of 1,000 per second, the peak value of emission from a well-activated oxide-coated cathode may be ten to one hundred times larger than the figure given above.

In designing an indirectly heated cathode for general-purpose and receiving tubes, it is generally desirable to make the ratio of length to diameter as great as possible, since end loss is then low and clearance between cathode and the next nearest electrode is maximum. On the other hand, provision must be made for insertion of the heater, and the length-to-diameter ratio must not be so great that bending or bowing of the cathode is likely to be serious. In practice, the ratio is generally about 25.

8. Flat or Ribbon Filaments. Where a cathode with a flat or nearly flat surface is required, such as in a cathode-ray tube or others in which the electron stream is formed into a cylindrical beam, the heater-power efficiency is low compared with that of a cylinder coated on the side, although this is frequently of minor interest. Where a new design is required, life tests are commonly made over a range of heater power, and the heater current rating is selected on the basis of life test results.

For the directly heated cathode, the design of a filament is made on the basis of performance records of tubes using the same material. Knowing the dimensions of the filaments of these tubes and the voltage and current for satisfactory life, the values of ρ and η are computed. The new filament, if round, may be computed as for the tungsten wire; corrections for the cooling effect of the leads and hooks are applied in the same manner.

The use of a ribbon filament provides an additional degree of freedom in choice of dimensions or voltage or current, since the surface area may be varied without changing the area of cross section.

The relations corresponding to Eqs. (6) and (7) for the ribbon filament will be

$$\rho = abE/IL \qquad \text{ohms} \tag{10}$$

$$\eta = EI/2(a + b)L \qquad \text{watts/sq cm} \tag{11}$$

where a and b are the width and thickness of the ribbon.

From the data on known or experimental tubes, the values of ρ and η are computed for the conditions that have been shown to give best results for life test. Then when values of length, current, and voltage are specified, the other dimensions will be

$$a = \frac{EI}{4\eta L} + \frac{1}{2}\sqrt{\left(\frac{EI}{2\eta L}\right)^2 - \frac{4\rho LI}{E}} \qquad \text{cm} \tag{12}$$

$$b = \frac{EI}{4\eta L} - \frac{1}{2}\sqrt{\left(\frac{EI}{2\eta L}\right)^2 - \frac{4\rho LI}{E}} \qquad \text{cm} \tag{13}$$

Since the thickness b is important and is the difference between two quantities apt to be nearly alike, computation should be carried out by the use of log tables in which the values of the logarithms are given to at least six digits. The result is conveniently

checked by substitution in Eq. (9). A negative or complex solution indicates the values chosen for E or I are too small, or that L is too great. In the practical case, the ratio a/b is desirably greater than 3 and less than 15.

Since ribbon filament is ordinarily made by rolling a round wire, the thickness is apt to be somewhat variable across the width. For this reason, ribbon filaments are ordinarily specified in terms of width (mils) and weight (mg/200 mm) to ensure against wide variations in current and temperature.

9. Preparation of Cathodes. The base material, whether it is wire, ribbon, or an indirectly heated cathode, is thoroughly cleaned by solvent materials such as dichlorethylene, then heated in a hydrogen atmosphere to a temperature of about 1000°C before coating. The coating is usually applied in one of three ways:

Bench coating consists in passing the wire or ribbon alternately over revolving wheels in a cup of coating, then through an electric furnace. The coating is thus built up in several layers. The carbonates are suspended in a solution of barium nitrate, the nitrate serving as a temporary binder. The furnace is supplied with a stream of carbon dioxide.

Spray coating consists in spraying the coating suspension on the ribbon or cathode with a spray gun, nitrocellulose being used as a binder. The desired texture of the coating is achieved through adjustment of nozzle, air pressure, and distance between nozzle and cathode.

Electrophoresis consists in applying the coating to the cathode by passing the wire through, or dipping the cathode into, a suspension of the coating and applying a strong electric field between cathode and metallic container.

In coating the cathode, the active materials are in the form of carbonates which are broken down to the oxides during the exhaust process. These carbonates, usually of barium and strontium, are ordinarily prepared especially for the purpose. Many of the larger tube-manufacturing firms prepare their own carbonates, while others purchase them from chemical supply houses in a form especially adapted to this work. The barium and strontium carbonates—sometimes calcium carbonate is also a constituent—are precipitated together from a hot solution of the nitrate by the addition of a solution of ammonium carbonate, sodium carbonate, or sodium bicarbonate, then washed and calcined. The temperature and concentration of the solution are closely controlled to produce crystals of the proper form; these crystals ordinarily have a diameter of a few microns.

10. Coating Suspensions. The composition of the suspension will depend upon the method of coating, the size of the wire, and the intended use of the cathode; the experience of the process engineer is also an influencing factor. For this reason, the coating formulas given below merely illustrate the way in which such suspensions are compounded; they are, however, practical for use in making experimental tubes.

1. Bench-coating formula:
   - 1,000 g carbonate
   - 1,000 cc distilled water
   - 100 g barium nitrate (decrease for wide ribbon)
2. For use in spray gun:
   - 575 g carbonate
   - 1,000 cc amylacetate
   - 20 g nitrocotton
3. Electrophoretic suspension:
   - 50 g carbonate
   - 5 g barium nitrate
   - 1,000 cc amyl acetate
   - 63 g nitrocotton

In all cases the suspension is ball-milled for a minimum of 12 hr to ensure complete separation and wetting of the individual particles. In using formula 1, it is necessary to heat the wire to a temperature high enough to fuse the nitrate or to convert it to

the carbonate. In the other cases, care is taken to ensure thorough evaporation of the amylacetate or other solvents. Plasticizers are frequently added in small proportions to give a degree of flexibility to the coating when nitrocotton is used as the binder.

11. Exhaust and Aging of Coated Cathodes. During the exhausting of the tube, the carbonates are broken down to the oxides by the application of heat, the reaction being summarized by the chemical formula

$$BaCO_3 \rightarrow BaO + CO_2$$

At atmospheric pressure the reaction does not proceed very far owing to equilibrium between the carbonate, oxide, and carbon dioxide; at exhaust the carbon dioxide is continuously removed, permitting the reaction to go to complete breakdown of the carbonate.

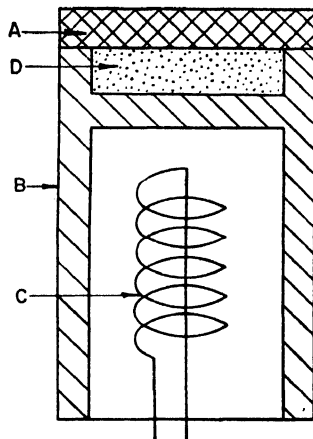
After the tube has been sealed off, it is ordinarily aged. The cathode is heated by the application of a heater voltage higher than normal. This may be as great as two to three times normal operating value, followed by a second step in which the heater voltage is held at about 150 per cent of normal with the cathode passing a thermionic current of about 100 ma per sq cm. The emission current is usually limited by an incandescent lamp ballast in the plate circuit. The time required for aging may be a few minutes for the first step and 10 to 20 min for the second step. Aging schedules vary widely depending upon such factors as the type of tube, the exhaust schedule, coating texture, and the composition of the base material. Their specification is ordinarily the result of cut-and-try experimentation.

FIG. 2. L-type cathode.

12. Base Materials. The base or core material used in most coated cathodes is nickel. Usually the nickel contains impurities which serve to liberate free barium at the interface between base and coating by reacting with barium oxide. The barium, in trace amounts, is required for activation of the coating. For filamentary cathodes, alloying materials are also added to improve the mechanical properties of the wire or ribbon such as hot strength.

The impurities in nickel that serve to produce free barium by interaction are silicon, magnesium, titanium, aluminum, and sometimes one or more other elements. Silicon is a very active reducing agent, but it forms a highly insulating compound, barium orthosilicate, at the interface and this produces an adverse influence on the performance of the tube. Consequently, the silicon content is usually held at as low a percentage as possible.

Cathode nickel, or nickel, alloys are commonly classified as *active*, *normal*, or *passive* according to rapidity or ease with which activation of the coating takes place. There is no sharp line of demarcation between them. An active alloy commonly contains 1 per cent or more of reducing impurities; a normal alloy usually has less than 1 per cent, while the impurity content of a passive alloy is negligibly small. In aging, and during life, the impurities diffuse to the surface of the core where reaction with barium oxide takes place, and as the concentration of reducing impurities in the nickel approaches zero the electron emission usually drops below a useful value.

13. Dispenser Cathodes. A dispenser cathode consists essentially of an emitting surface supplied with an activating material from a reservoir, and a heater element as shown in Fig. 2. Best known is the "L" cathode, in which the active part of the cathode consists of a porous tungsten plate *A* covered by barium. The cylinder *B* into which all parts are assembled is commonly made of molybdenum, the heater *C* is alundum-covered tungsten wire, and the reservoir *D* is filled with (Ba, Sr)CO_3 or

the triple carbonate. When the tube is exhausted, the carbonate is reduced to the oxide.

Operation is, however, different from that of the ordinary cathode. The oxide in the reservoir is partially reduced with the liberation of a small amount of free barium; the latter migrates through the pores of the plate to the top surface. The work function of this cathode is usually about 1.60 to 1.65 electron volts, which is taken to mean that the surface of the tungsten is oxidized and almost completely covered by a monatomic layer of barium. Since the work function is higher than that of the ordinary coated cathode, dispenser cathodes must be operated at a higher temperature to obtain the same emission. The dispenser cathode is of special interest in devices requiring a high current density from a small area of the cathode, especially where ion bombardment is severe enough to ruin a coated cathode.

A modification of the dispenser cathode just described is the impregnated cathode, in which the reservoir *D* is omitted. Activating materials, commonly a mixture of normal and basic barium aluminates, or aluminates and carbonates, are deposited between the grains of the porous tungsten plate.

14. Matrix Cathodes. A matrix cathode is in some respects similar to the dispenser cathode in that the activating material is placed in, instead of on, the emitter. Powdered nickel is used as the base, and materials such as $(Ba,Sr)CO_3$ are compacted with the nickel. Alternatively, the nickel may be sintered to form a porous body, and the carbonates then precipitated within the interstices.

It is indicated by measurements of emission vs. temperature that a fully aged matrix cathode has a work function of about 1.05 to 1.10 electron volts; consequently it is assumed that the surface is covered with BaO or (Ba,Sr)O. Such cathodes have limited application and were still in the development stage when this book was published.

15. Testing Thermionic Emission. The thermionic-emission properties of a tube are measured in a number of different ways, the choice being determined by the type of tube or the purpose of the test. A number of these tests will be described briefly. In most cases all electrodes except cathode are tied together at the socket to form a composite anode, and the anode current is noted under conditions which may be specified.

Ideally, the (composite) anode current of a tube should increase with the three-halves power of the anode voltage, when the cathode temperature is constant, until temperature-limited emission is reached. From this point the current should increase only gradually as the field at the cathode is increased. In the actual case, especially with coated cathodes, it is often difficult or impossible to ascertain the point at which the transition from space-charge-limited to temperature-limited current occurs because of lack of uniformity in the temperature and emission constants of the cathode. Results are more meaningful if plotted as a Schottky line from which the true value of emission may usually be inferred.

The Schottky relation shows the manner in which the electron current is enhanced when the field is accelerating throughout the cathode-anode region. Here the total current I for ideal conditions is

$$I = I_0 \epsilon^{e\sqrt{-eF}/kT}$$

where F = electric field at cathode surface; I_0 = normal emission at $F = 0$; e = electron charge; k = Boltzmann's constant; T = temperature of cathode in absolute degrees.

This expression can be written $\ln I = \ln I_0 + e\sqrt{-eF}/kT$. When converted to common logarithms with the constants evaluated the result is

$$\log I = \log I_0 + (1.9/T)\sqrt{-F}$$

in which the current is expressed in amperes/cm² and the field in volts/cm. In most cases the emission is too high to test at normal cathode temperatures; hence pulse techniques are used. If the true temperature and field are not known, one may still

determine emission to a satisfactory degree of accuracy by taking the above expression in the form $\log I = \log I_0 + K\sqrt{e_p}$ in which K is a constant whose value need not be known and e_p is the plate voltage. If $\log I$ is plotted against a number of values of e_p the curve, at low values of plate voltage, is the three-halves-power relation; at high values of plate voltage the current increase, which is much less than in the space-charge region, is due mostly to the effects described by the Schottky relation. Therefore, if the curve in the Schottky region is extrapolated to $e_p = 0$, the intercept on the vertical axis gives the logarithm of the normal emission of the cathode.

For tubes having coated cathodes, care must be taken to avoid overheating. In many cases the emission may fall because of the decomposition of impurities on the anode so that materials poisonous to the cathode are released. For these reasons, approximation methods are frequently employed.

Some direct emission tests are as follows:

1. For small tubes having tungsten filaments, the emission current is measured at rated filament voltage and with the composite anode voltage at or near the value required to produce temperature-limited emission. Care must generally be exercised in keeping the time of test to a minimum to avoid overheating the electrodes.

2. For small tubes having coated filaments or coated, indirectly heated cathodes, the composite anode voltage is commonly set at a value which, in a well-activated tube, produces an emission current of approximately 200 ma per sq cm of emitter surface. The test must not be too long continued, although, for purposes of quality control, it is made for a long enough time–say 3 to 10 sec—to ensure that the emission is reasonably stable.

FIG. 3. Flection-point emission. Line a is hypothetical curve of space-charge current plotted against composite anode voltage. In the actual case the current is limited by emission line b. Tangent point x is taken as flection-point emission.

3. Where a very high degree of cathode activity is required and where the application of a voltage sufficiently high to result in a current near the temperature-limited value would result in serious injury to the tube, the anode voltage is pulsed at a relatively low duty cycle.

4. For many purposes, the "flection-point" emission is taken as a criterion of cathode activity. The flection point is defined as the point on the diode characteristic (current plotted against composite anode voltage) at which the second derivative of current with respect to voltage reaches its maximum negative value. This point is often referred to as the "knee of the curve." Although not rigorous, it is conveniently taken as point x, Fig. 3, the point of tangency between the diode characteristic and a straight line through the origin. Data may be taken point by point or by means of a cathode-ray oscillograph.

Indirect checks depend for their value upon the increase in emission with temperature, the total emission or tube performance being measured at a cathode temperature lower than normal. Two such checks will be described.

1. In tubes having filamentary cathodes that might be injured by the passage of a relatively high emission current, the check consists in measuring the filament voltage required for a substantially temperature-limited emission current of a specified

magnitude. For such a test the current will ordinarily be approximately that of the total cathode current under ordinary operating conditions.

2. For large transmitting tubes, an indirect emission check consists in testing the power output of the tube operating as a self-excited oscillator with standard voltages, then reducing the filament voltage until the power output is decreased to a specified fraction of the original value. The filament voltage reading under this condition then constitutes an indirect check upon emission. The value of such a check depends, to a considerable degree, upon judicious choice of circuit conditions and is applicable only in comparing individual tubes of the same type.

The cathode heating time is related to emission testing and will be mentioned here. Conventionally, this is the time which elapses between application of rated filament or heater voltage and the attainment of a cathode current such that its time rate of change is maximum, corresponding approximately to point x in Fig. 3. The primary of a transformer is placed in the cathode circuit with its secondary shunted by a microammeter; maximum rate of change occurs when the reading of the microammeter reaches maximum.

16. Photoelectric Emission. The energies of the conduction electrons of a solid material may be raised to a magnitude sufficiently high to permit their emergence from the body by means of radiation without heating. When emission occurs through the agency of radiation, usually within or near visible wavelengths, the emission is classified as *photoelectric emission*, the emitter is a *photocathode*, and the electron tube, whether gas-filled or vacuum, employing a photocathode is a *phototube*.

For what one might loosely call "white light," the magnitude of the photoelectric current for unit emitting area is proportional to the intensity of the light falling upon the photocathode and depends upon the work function of the surface in much the same manner as in the case of thermionic emission. On the other hand, the magnitude of the current also depends upon the color composition of the incident light for the reason that the energy which the light may give to an electron is inversely proportional to its wavelength. For a given emitter there will be a maximum wavelength, called the *threshold wavelength*, at which photoelectric emission is obtained; for shorter wavelengths the excess energy imparted to the escaping electron beyond that required to overcome the surface forces will be the kinetic energy the electron will have upon emergence. Quantitatively stated,

$$hc/\lambda = \phi e + mv_0^2/2 = h\nu \qquad (14)$$

where h = Planck's constant
c = velocity of light
λ = wavelength of radiation
ϕ = work function of photocathode, volts
e = electron charge
m = mass of electron
v_0 = velocity of electron upon emergence[1]
ν = wave frequency = c/λ in air or vacuum

The threshold wavelength is, therefore, the wavelength at which the second term of Eq. (14) just becomes zero. When the constants are evaluated in practical units and the work function ϕ is expressed in volts, the threshold wavelength λ_0 is given by

$$\lambda_0 = 12{,}395/\phi \quad \text{A} \qquad (15)$$

where the wavelength is measured in angstrom units, A = 10^{-8} cm.

As an example, if a photoelectric current is to be obtained which is sensitive to all the visible colors, and the limit of visibility is taken as 7000 A at the red end of the spectrum, the work function of the photocathode must be not greater than 12,395/7,000 = 1.77 volts. Many phototubes have their long-wave limit extending well into the infrared region. These tubes are not, however, to be confused with the

[1] The velocity v_0 will be greater by a small amount when the energy before excitation is added to both sides of the equation.

infrared cells such as the thallous sulfide type which are sensitive at wavelengths from within the visible range to and beyond 12,000 A.

While Eqs. (14) and (15) are intended to be rigorous only at absolute zero, the result is not greatly different at room temperature for the reason that the chance that a given electron shall be excited by the radiant energy appears to be independent of the energy it already had. Reference to Fig. 1 shows that the average energy is not greatly altered as the temperature is raised.

Most photoelectric cathodes exhibit a spectral selectivity such that the maximum response at a given level of incident radiant energy is obtained at a wavelength of or fairly near $2\lambda_0/3$. In the case of thin composite surfaces, such as those comprising an oxidized silver base upon which an alkali metal such as cesium or rubidium is deposited, two and sometimes three maxima are found, each being characteristic of an element or compound subjected to the radiation.

The "complete" photoelectric emission is expressed by an equation identical in form with that for thermionic emission. If the photoelectric cathode is exposed to the total thermal radiation from a black body at temperature T, the total photoelectric current for clean metals varies with T in accordance with the relation

$$I = A_1 T^2 \epsilon^{-h\nu_0/kT} = A_1 T^2 \epsilon^{-b_0/T} \quad \text{amp/sq cm} \tag{16}$$

in which A_1 is found to have values usually of about 10^{-10} to 10^{-9} amp per sq cm per deg^2 and $\nu_0 = c/\lambda_0$. The numerical value of b_0 is $1.16 \times 10^4 \phi$ deg as in the case for thermionic emission. This could be expected since the work ϕe done by the electron in leaving the cathode is the same regardless of whether its energy within the body is of thermal or photonic origin.

17. Photocathodes. The choice of a photoelectric tube depends, naturally, upon the use for which it is intended. The best guides in the choice of tube are the catalogues or handbooks of the most reliable manufacturers of such devices. The brief description to be given of some kinds of phototube cathodes is intended only for the general reader.

Of rather specialized interest are solid metal cathodes or cathodes coated to a relatively great thickness—*e.g.*, 100 or more atomic diameters—with metals having relatively high work function, such as zirconium, tantalum, or thorium. These are of interest in the measurement of ultraviolet light and usually are mounted in bulbs transparent to the desired regions of ultraviolet or employ quartz windows. Chemical cleanliness of the materials is of greatest importance; proper bulb baking and outgassing by induction heating are required for reasonably reliable results.

The sensitivity of some of these surfaces may be considerably enhanced and the long-wave limit λ_0 raised by passing a glow discharge between cathode and anode in the presence of hydrogen. This is preferably done after outgassing, the hydrogen being then pumped out before the tube is sealed off. Gases other than hydrogen, *e.g.*, sulfur vapor, have been tried with success.

Of greater interest are the photocathodes produced by the deposition of thin films of certain alkali metals upon a solid metal support. Because of the low work function of these surfaces, some of them are sensitive in the visible regions of the spectrum. A summary of features of present interest concerning some of these surfaces is shown below, the values given being approximate.

Surface	λ_m, A	λ_0, A
S-1 (commonly Co-O-Ag)	8,000	12,000
S-3 (commonly Rb-O-Ag)	4,300	9,000
S-4 (commonly Cs-Sb)	4,000	7,000
S-5 (commonly Cs-Sb)*	3,400	7,000
S-8 (commonly Cs-Bi)	4,200	8,000

* In ultraviolet transmission bulb.

λ_m is wavelength for maximum sensitivity; λ_0 is long-wavelength limit.

Among the most useful commercial photocathodes, from the standpoint of sensitivity, is the Cs-CsO-Ag emitter. Although the maximum sensitivity of this surface does not correspond to that of the human eye—none of them do—its performance under tungsten or daylight illumination is of considerable use in commercial devices. It consists essentially of a well-cleaned silver support which, during exhaust, is outgassed, then oxidized by the passage of a glow discharge in oxygen. Care must be exercised in obtaining the proper depth of oxidation, the progress being observed by watching the subtractive interference color changes. The uncombined oxygen is then removed and cesium distilled into the bulb, care being taken to admit only as much cesium as is required to produce the proper sensitivity.

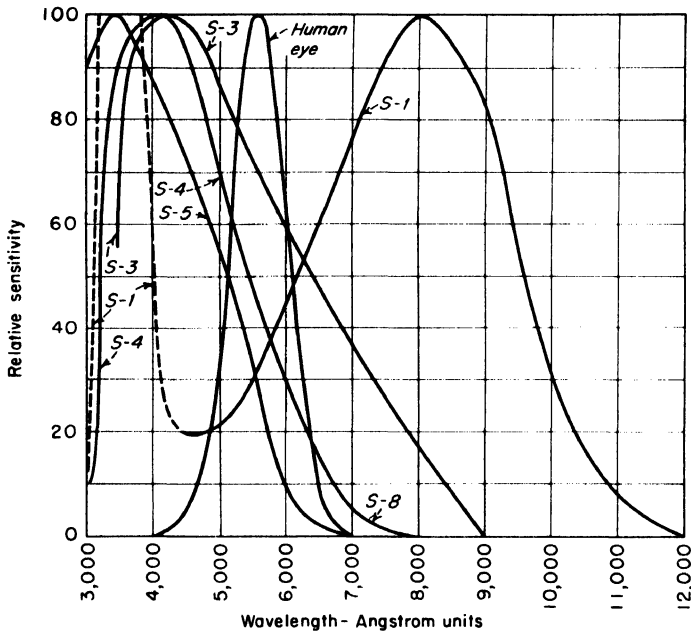



FIG. 4. Average response of commercial photocathode surfaces compared with that of the average human eye.

The response characteristics of some commercial phototube surfaces are shown in Fig. 4, wherein the average spectral response is compared with that of the average human eye. Relative sensitivities for each curve are plotted against wavelength for equal values of radiant flux at all wavelengths at the photosurface except for the S-5 photocathode, which is plotted for constant flux outside the bulb.

18. Multiplier Phototubes. A multiplier phototube comprises essentially a photocathode, an anode, and one or more dynodes[1] sealed in an evacuated transparent envelope or an envelope having a transparent window. Commercial multiplier phototubes commonly have a sensitivity of 5 to 10 μa per microlumen. Cesium-bismuth and cesium-antimony photocathodes are commonly used.

In operation, the first dynode is maintained at a potential positive with respect to the photocathode, and each succeeding dynode is positive with respect to that preceding. Electrons photoelectrically emitted from the cathode are drawn to the first dynode; with a secondary emission ratio greater than unity, secondary electrons are emitted which, in turn, give rise to an amplified current from the next dynode. By the time the electron stream has reached the anode, the original photocurrent is amplified

[1] Secondary electron-emitting electrode.

δ^n-fold, *where δ is the secondary electron emission ratio of the dynode and n is the* number of dynodes.

The multiplier phototube has a number of advantages compared with a phototube and separate amplifier. Since the amplification is entirely that of a single primary stream of electrons, the incident light which actuates the current may be relatively constant or it may be chopped, as by means of a rotating shutter, and the gain remains constant. Of still greater advantage, especially for the measurement of light at low intensity, is the high signal-to-noise ratio obtained.

In considering the signal-to-noise ratio of the multiplier phototube, it should be noted that total absence of light does not result in zero anode current. The small current observed under such conditions is known as the "dark current," the principal components of which are (1) amplified thermionic emission, (2) leakage current, (3) regenerative ionization. At high levels of cathode illumination, the dark current is likely to be of minor importance; at low levels, such as in astronomical work, the dark current may be important, although the resultant "noise" is still low compared to what would be obtained in a unit containing a phototube and separate amplifier having the same gain and a band width equal to that of the equipment into which the multiplier phototube operates.

19. Secondary Emission. A third way in which electron emission may be produced is by the bombardment of a conducting or semiconducting solid with electrons or positive ions, known as *secondary emission.* Such an emitter, when used as an electrode of an electron tube, is called a *dynode.*

Secondary emission may be obtained by electron bombardment of pure metals, although the yield is not great. In this case a part of the energy of the bombarding or *primary electron* is transferred to one or several conduction electrons of the solid. If the surface-directed energies of the newly excited electrons are greater than the work function ϕe, the energy required for their exit, they may be emitted as *secondary* electrons. In the case of pure metals, the ratio of secondary electron current to primary electron current, δ, is not appreciably affected by temperature.

When the secondary-electron-emitting electrode or dynode is a composite surface properly prepared for use as such, δ may be made considerably greater than unity with the result that a weak electron current, such as that from a photocathode, may be amplified many times by secondary emission. Other uses of the phenomenon are found in the dynatron and in the orbital-beam tube. The secondary emission ratio of such dynodes is not entirely independent of temperature.

The mechanism of secondary emission from composite surfaces is generally considered to be different from that of pure metals. In the latter case the "free" electrons of the metal are thought to provide the secondary electron current. In the case of composite surfaces, certain compounds of the metals, especially those of the second and third columns of the periodic table, provide much higher values of δ than do the pure metals. Among the best compounds are barium oxide and magnesium oxide; sulfides, halides, etc., of these metals are not satisfactory. Here the secondary emission is provided by the excitation, due to the energy of the incident primary electron, of the electrons of the outermost shells of atoms constituting the emitting material. In such a case it is considered that the permitted energy bands for electrons in the compound are either entirely empty or completely filled, and that a bombarding primary electron can lose its energy by causing the transition of an electron from a filled band to an empty band. The chance that an excited electron shall leave the body depends upon the position of the empty band with respect to the occupied band *and* the potential outside the body.

If all the energy brought into the body by the bombarding primary could be transformed into the energy required for the emission of secondary electrons, the ratio δ would be very high. If V is the voltage between dynode and cathode, the energy of the primary just inside the dynode would be $Ve + \phi e$; the energy would be transformed to n secondary electrons having a total energy $n\phi e$. Thus the maximum number n that could be emitted would be $n = V/\phi + 1$. There are a number of reasons why this is not realized in the actual case:

1. The energies of the excited electrons are randomly distributed in direction; it is only the surface-directed component of the energy of an individual electron that determines whether it shall be emitted.

2. The majority of the emitted secondary electrons receive considerably more energy from the primaries than is required for emergence. Thus the average initial velocity of secondary electrons is greater than in the case of thermionic or photoelectric emission.

3. The energy of the primary electron, especially when the potential of the dynode with respect to the cathode is high, may not be expended until the primary electron has penetrated to a considerable distance below the surface of the dynode, in which case the chance that electrons excited by the primary shall be emitted is considerably reduced.

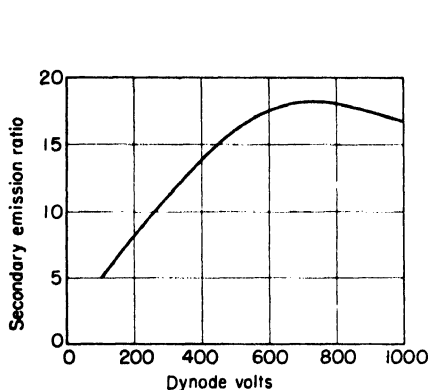



FIG. 5. When primary electron stream is directed normally to the surface of the dynode, the secondary emission ratio reaches a maximum value as the primary energy is increased. Curve is for magnesium oxide. (*After Bruining and de Boer.*)

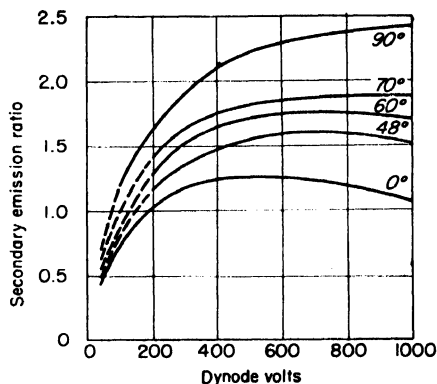



FIG. 6. Secondary emission ratio at various angles of incidence of primary electrons. Curve is for nickel. (*After Bruining and de Boer.*)

Typical examples of the manner in which the depth of penetration of the primary electrons affects the magnitude of the secondary electron current are shown in Figs. 5 and 6, both of which were redrawn from the data of Bruining and de Boer.

In Fig. 5 the ratio δ of secondary to primary current for magnesium oxide–coated platinum dynode is seen to rise from zero at very low dynode voltage almost linearly at low voltage, then gradually reaches a maximum value at about 700 volts, after which δ falls to lower values as the voltage is further increased. The curve is taken as a demonstration of the contention that, when the energy of the primary electron is high, much of it is transferred to other electrons at a distance so far below the surface that the chance of a secondary electron being emitted is greatly reduced.

Figure 6 shows the manner in which δ varies with the angle of incidence of the primary beam when the magnitude of the primary current and the potential of the dynode are held constant. It is assumed that, as the angle between the beam and the surface of the dynode is reduced, the depth of penetration of the primary electrons is reduced; hence the average distance between the surface and the point at which the maximum excitation occurs is lowered. The result is an increase in the secondary electron current or δ.

In practice, dynodes are made by activating the surface of a base metal either by oxidation, by distillation onto it of a sensitizing material, or by heat-treatment to bring an impurity sensitizer to the surface.

An example of the first-named class of activation is that in which the base metal is

an alloy consisting largely of magnesium, such as Dowmetal. The dynode is cleaned by mechanical processes, then heated in an atmosphere of oxygen from which water vapor and nitrogen have been excluded to produce a thin film of magnesium oxide.

In sensitizing by the distillation of sensitizing materials onto a base, one of several processes may be used. Barium or magnesium, or a combination of both, may be volatilized onto a nickel base *in vacuo*, then oxidized by admitting a small amount of oxygen to the device and applying a spark coil so that a glow discharge is created.

A useful method comprises using a well-cleaned and outgassed silver or copper electrode upon which a film of cesium is deposited by vaporization from a side tube; the cesium is then oxidized by a glow discharge in low-pressure oxygen. After removal of the residual oxygen, a further small quantity of cesium is admitted. Dynodes made by this process have superior properties when properly prepared, the reason apparently being that the secondary electrons originate within the cesium oxide lattice where they are readily excited, while the metal cesium coating, because of its low work function, facilitates their escape.

For activation by means of an impurity material in the dynode, the dynode ordinarily consists of silver containing a few per cent by weight of magnesium. Activation consists in heating the dynode for a period of at least several minutes in the presence of oxygen, during which time some of the magnesium diffuses to the surface where it becomes oxidized.

In most cases where secondary emission is used, care is usually taken to ensure that the dynode is shielded from being contaminated by the cathode. While it is true that an electrode may be activated solely by contamination from an oxide-coated cathode, such activity is low compared to that produced by the methods just described and has relatively short life. A good dynode is, therefore, effectively spoiled by cathode contamination. The orbital-beam tube is an example of one way in which the dynode is protected.

Secondary emission often occurs as a spurious effect in various kinds of electron tubes, especially through transfer of activating material from the cathode to the anode or a grid. This is particularly noted in tubes employing oxide-coated cathodes. In the case of tetrodes the effect appears as a major shift in screen-grid current and in plate resistance during the life of the tube even with no change in cathode current. A fairly effective remedy for this condition is to spray the grid with carbon black or boron carbide and to carbonize the plate by first oxidizing the surface, then heating it in a carbon-bearing gas such as methane, CH_4. The real solution to the problem is, of course, the employment of a suppressor grid held at a low voltage such that the field on both sides of it suppresses the emission of secondary electrons.

Secondary emission plays an important part in the operation of television pickup tubes; it is beyond the scope of this book to describe such a highly specialized device in detail. Secondary emission is also important in cathode-ray tubes in that it constitutes the means whereby the electrons striking the screen are able to reach the terminal for the screen voltage supply.

20. Field Emission. From an engineering standpoint, field emission is of less interest in the radio electron tube than are thermionic, photoelectric, and secondary emission currents; hence the discussion will be confined to a few remarks.

The relations describing quantitatively the phenomena of thermionic and photoelectric emission assume zero field at the surface of the cathode. When the field is not zero, the effect is equivalent to a change in the net work function by an amount

$$\Delta\phi = -\sqrt{-eF} \tag{17}$$

where F is the field or potential gradient just outside the surface of the cathode. In practical units,

$$\Delta\phi = -3.8 \times 10^{-4} \sqrt{-F} \qquad \text{volts} \tag{18}$$

in which F is in volts per cm.

Field emission is of greater interest in gaseous discharge devices than in high-vacuum tubes, but the phenomenon is one which must be considered in any application where strong fields are encountered, as in transmitting and rectifier tubes.

21. Spurious Emission Currents. In many cases the operation of a tube may be adversely affected by unwanted emission currents, and special means must be taken to prevent or reduce them. A brief description of the most common practices will be given.[1]

Grid emission may occur by any of the means already mentioned, the most common being thermionic or primary emission, and secondary emission; photoelectric emission is encountered in some tubes where the grid current due to gas or leakage is low enough for the photoelectric effect to become a factor.

Primary grid emission is encountered in receiving tubes employing oxide-coated cathodes where the spacing between grid and cathode is low and the dissipation on the electrodes is high in relation to their area. This is particularly noted in the output tubes, especially those intended for use in a-c–d-c receivers.

In most of the commercial receiving tubes where coated cathodes are used and where, without special precautions, grid emission would be a factor, radiators are welded to the ends of the grid side rods to keep the grid cool. Attention is given to the thermal conductivity both of the laterals and of the side rods; pure metals are preferred for the former, while beryllium-hardened copper is frequently used for the latter. For tubes used in oscillator service, the control grid is sometimes gold-plated or sprayed with finely powdered boron carbide if the type of operation is sufficiently severe to warrant this added expense.

For tubes employing thoriated tungsten filaments, the means employed to hold the grid emission to a low value depend to some extent upon the anticipated maximum grid temperatures. Where the temperature is not excessive, the grid may be made of platinum or of a more refractory wire which is platinum-clad, since platinum has the ability to absorb any thorium deposited upon it. For extremely high temperatures, the grid may be coated with certain refractory oxides, provided they do not break down with evolution of oxygen and cause decarburization of the filament. Good results have been reported with the oxides of zirconium and beryllium on molybdenum grids.

The means employed for the reduction of primary grid emission are usually effective in the control of secondary emission. However, the latter effect may be encountered in tubes having grids that operate at relatively low temperatures, in which case they may be sprayed with substances having a low coefficient of secondary emission such as carbon black, graphite, or zirconium.

Tubes are frequently designed for use wherein the anode may, during part of a cycle, be at a potential lower than some other electrode, as in rectifiers and tetrodes; hence either primary or secondary emission from the anode may be a serious factor unless precautions are taken to prevent or limit its occurrence. Where the tube is a diode rectifier, emphasis is placed upon the radiating properties of the anode so that the temperature will be low enough to limit primary emission. For small tubes, the anodes are commonly made of carburized sheet nickel. In large tubes the anode is made with radiating fins and is sand-blasted and coated with zirconium, the latter serving both as a getter and in improving thermal radiation.

For tetrodes, the problem of thermionic emission from the anode is less frequently encountered. Secondary emission may be held at a minimum by the use of carburized metal, or by a zirconium spray on the inside of the anode. To a considerable extent, however, this problem is more of academic than of practical importance, since the use of a suppressor grid as used in pentodes is a more practical solution.

SPACE-CHARGE-LIMITED CURRENTS

22. Diodes. A space-charge-limited current in an electron tube is an electron current which depends only upon the configuration and potentials of the electrodes

[1] An excellent discussion of the effects of operating conditions on spurious currents and on changes in electron emission is given in "Techniques for Application of Electron Tubes in Military Equipment," Document PB111644, U.S. Department of Commerce, Office of Technical Services, Washington 25, D.C.; price $2.50.

and is *not* limited by the thermionic emission of the cathode. In this section all the relations are to be taken as being of practicable value only to the extent that adequate emission is available.

In a diode having a unipotential cathode, the current is given by the relation

$$i = Ge_p^{3/2} \tag{19}$$

in which G is the perveance and e_p the instantaneous anode voltage. In the formulas which follow, G will be expressed in terms of amperes per volt$^{3/2}$.

Where the cathode is a filament, the voltage drop E_f across the filament must be considered, and this requires modification of Eq. (19). Commonly the anode voltage is larger than the filament voltage, and return connections are made to the negative end of the filament. For this case the current is expressed as

$$i = \frac{2}{5}\frac{G}{E_f}\left[e_p^{5/2} - (e_p - E_f)^{5/2}\right] \tag{19a}$$

where G is now the perveance for a unipotential cathode tube of the same dimensions.

Where the value of e_p is several times that of E_f, it is ordinarily of sufficient accuracy to use the relation

$$i = G(e_p - E_f/2)^{3/2} \tag{19b}$$

for connection to the negative end of the filament. For connection to the center tap of a filament-heating transformer, E_f in Eq. (19*b*) is omitted in considering the d-c component of the current.

For *parallel-plane* electrodes,[1]

$$G = 2.34 \times 10^{-6} A/x^2 \tag{20}$$

where A = electrode area
x = electrode spacing

The physical dimensions may be taken in any units for Eq. (20) or the formulas to follow, since the units cancel out.

For *coaxial cylinders*,[1]

$$G = 14.65 \times 10^{-6} L/r\beta^2 \tag{21}$$

where L = length of electrodes
r = radius of anode
β^2 = a function of r/r_o shown graphically in Fig. 7, r_o being the cathode radius

For tube electrode configurations other than parallel planes and coaxial cylinders approximation solutions are employed in computing the perveance. These are based on the equivalent-capacitance rule, which states that the space-charge-limited current between a cylindrical cathode and an anode of arbitrary shape is approximately that which would be expected for the same cathode and a symmetrical anode of such dimensions that the electrostatic capacitances are equal. The current relations for most of the electrode arrangements used in tubes may be derived from one or another of the seven expressions for the capacitance given below. In all these cases the capacitance is to be equated to that for a pair of coaxial cylinders, the radius of the outer cylinder being r and that of the inner one r_o. The resulting expression is then solved for r, and this value is then used in Eq. (21).

[1] The perveance equations (20) and (21) do not include corrections for initial electron velocity and contact potential difference; consequently they are only approximations for small spacings and low voltages. For parallel planes, the distance s is to be reduced by s_{min}, the distance between cathode and plane of potential minimum or virtual cathode. Where $i_s \gg i$, it is sufficiently accurate in most cases to take this distance as $s_m = 2.7 \times 10^{-8} T^{3/4}/i^{1/2}$ cm, where T is the temperature of the cathode in deg K and i is the current density in amp per sq cm.

In the case of coaxial cylinders where r/r_o is not great, the correction in r_o is somewhat less than for planes and becomes vanishingly small for high values of r/r_o.

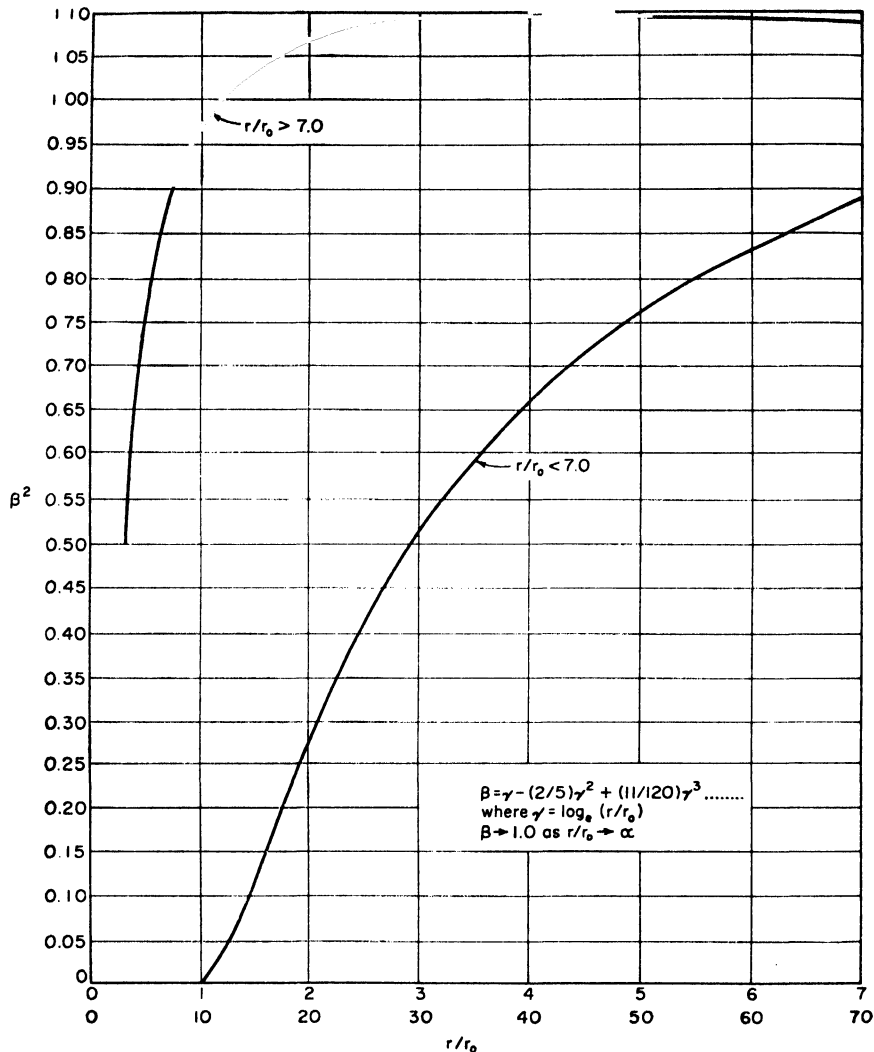



FIG. 7. Values of β^2 for r/r_o.

For coaxial cylinders, the capacitance is

$$C = \frac{l}{2 \ln (r/r_o)} \tag{22}$$

where l = length.

As an example of application, the rule will be applied to the case for a cylindrical anode surrounding a cylindrical cathode, the axes being separated a distance a. The capacitance for this electrode arrangement is

$$C = \frac{l}{2 \cosh^{-1} [(r_a^2 + r_o^2 - a^2)/2r_a r_o]} \tag{23}$$

where r_a = anode radius
r_o = cathode radius

When this expression is equated to Eq. (22), one has

$$\ln (r/r_o) = \cosh^{-1}\left(\frac{r_a^2 + r_o^2 - a^2}{2r_a r_o}\right) \tag{24}$$

To solve for the current, one simply takes

$$\frac{r}{r_o} = \frac{1}{2.3} \log_{10}^{-1}\left[\cosh^{-1}\left(\frac{r_a^2 + r_o^2 - a^2}{2r_a r_o}\right)\right]$$

to obtain r and r/r_o. This value of r is substituted in Eq. (21) along with the proper value of β^2 obtained from Fig. 7.

The expressions for the remaining six arrangements of practical importance are as follows.

Cylinder Parallel to a Single Plane:

$$C = \frac{l}{2 \cosh^{-1} (h/r_o)} \tag{25}$$

in which h is the distance from cathode axis to plane. It should be noted that, when $h \gg r_o$, the equivalent cylindrical anode has a radius $r = 2h$.

Cylinder Midway between Parallel Planes:

$$C = \frac{l}{2 \ln [(4h^2 - r_o^2/2)/\pi h r_o]} \tag{26}$$

where h = distance from cathode axis to one plane.

For the case in which $h \gg r_o$, the expression may be simplified and one has

$$r = \frac{1.27h}{r_o}$$

Two Cylindrical Cathodes in a Cylindrical Anode:

$$C = \frac{l}{n\left\{\left[\dfrac{r_a^2 - a^2}{(r_a^4/a^2) - r_o^2}\right]\left[\dfrac{r_a^4/a^2 - (r_o + a)^2}{r_o^2 + 2r_o a}\right]\right\}} \tag{27}$$

In this expression it is assumed that both cathodes are on a line with the cathode axis, the cathode axes being displaced a distance a from the anode axis, and that $r_a \gg (r_o + a)$. The length of the equivalent coaxial cylinders is taken as twice the length of the two-cathode structure.

Cylindrical Anode with n Cylindrical Cathodes:

$$C = \frac{nl}{2n \ln (r_a/a) + 2 \ln [(a/nr_o)(1 - a^2/r_o^2)]} \tag{28}$$

In applying the rule, Eq. (28) is to be equated to

$$C = \frac{nl}{2 \ln (r/r_o)}$$

Cylindrical Anode with Helical Cathode:

$$C = \frac{L}{(2L/l) \ln (r_a/a) + 2 \ln [(al/Lr_o)(1 - a^2/r_o^2)]} \tag{29}$$

where l = length of helix
L = total length of wire in helix
a = radius of helix
r_a = radius of anode
r_o = radius of filament wire
In applying the rule Eq. (29) is equated to

$$C = \frac{L}{2 \ln (r/r_o)}$$

Folded Filaments. In tubes using wire or small ribbon filaments folded as an M or a V and placed between rectangular grids or anodes, the capacitance is computed as though a single wire were used in a system whose length is the emitting length of the filament. Even with ribbon filaments, the spacing is usually large compared with the dimensions of the cathode so that the simple relation $r = 1.27h$ is satisfactory for use in predicting the current.

23. Contact Potentials and Other Internal Effects. Where e_p of Eq. (19) is of the order of 10 volts or more, this can be taken as the instantaneous voltage e_b applied at the terminals of the tube. At lower voltages, e_p must be taken to include certain internal effects, primarily the contact potential difference between the electrode surfaces and initial electron velocity. As the voltage becomes very low, the current departs from the three-halves law of Eq. (19) and is said to be limited by a retarding field; usually some current is still obtained even when the applied voltage is slightly negative.

The contact potential difference between the electrodes is taken as minus the difference in the work functions ϕ_k and ϕ_a of their surfaces. Thus if there is no potential difference applied between the two electrodes, the potential *just outside* the anode is $-\phi_a$ volts with respect to the inside of either electrode, and the potential *just outside* the cathode is $-\phi_k$. So far as the field is concerned, it is the same as though the anode is placed at a potential $-(\phi_a - \phi_k)$ volts with respect to the cathode.

The principal effect of the initial electron velocity is the existence of the so-called *virtual cathode* at some distance from the cathode, the potential of which is less than any other place between the electrodes. For parallel planes, the potential of the virtual cathode with respect to the outside of the cathode is given as

$$V_{\min} = (-T/11{,}600) \ln (i_s/i) \tag{30}$$

where T = temperature of cathode, degrees K
i = current to anode
i_s = temperature-limited emission

While the contact potential difference and the voltage $V_{\min}$ are not the only internal correction voltages, together they constitute the important ones from an engineering standpoint, their sum being denoted by the symbol ϵ. Thus Eq. (19) may be taken as

$$i = Ge_p^{3/2} = G(e_b + \epsilon)^{3/2} \simeq G[e_b - V_{\min} + (\phi_a - \phi_k)]^{3/2} \tag{31}$$

where e_b is the externally applied voltage and ϵ is the internal correction voltage, its value ordinarily being of the order of ½ volt for small tubes when well aged.

Figure 8 shows graphically the manner in which ϕ_k, ϕ_a, $V_{\min}$, and e_b combine to control the space-charge-limited current in a tube. Taking the potential inside the cathode as zero, the potential just outside the cathode will be $-\phi_k$ volts, depending upon the material of which it is made. When the cathode is at a temperature high enough to supply a space-charge-limited current with an applied voltage e_b,

the potential at all points is as shown by curve *a*. Owing to initial velocity, more electrons are emitted than reach the anode; their charge results in a depression in the curve at point x on curve *a*. The potential at point x is $V_{\min}$ with respect to the outside of the cathode, or $(-\phi_k + V_{\min})$ with respect to the inside of the cathode. Now the *inside* of the anode is at a potential e_b with respect to the *inside* of the cathode; therefore, the *outside* of the anode is at a potential $(e_b - \phi_a)$ with respect to the inside of the cathode. The potential difference represented by e_p is here taken as that between the outside of the anode and the virtual cathode; hence

$$e_p = (e_b - \phi_a) - (0 - \phi_k + V_{\min})$$

which is the term within the brackets of Eq. (31).

Line *b* of Fig. 8 shows the conditions existing when the cathode is cold; the potential minimum is zero; hence $V_{\min} = 0$ and $e_p = e_b + \phi_k - \phi_a$.

Curve *c* is for the case where the cathode is heated and $e_b = 0$; the potential difference between anode and cathode is $\phi_k - \phi_a$. Since i is now less than when $e_b > 0$, $V_{\min}$ (point x') is a greater negative voltage than before and is located much nearer to the anode.

FIG. 8. Effects of work functions and potential minimum on the field in a diode with parallel plane electrodes. Curve *a* is the potential distribution for a space-charge-limited current; curve *b*, potential distribution when emission is very low; curve *c*, potential distribution with cathode hot but with zero applied voltage.

The experimental value of ϵ may be determined to a useful degree of accuracy when the plate current and plate resistance are known. If it is assumed that the perveance G and potential minimum $V_{\min}$ in Eq. (31) are relatively constant with change in applied voltage e_b, the derivative de_b/di may be taken as the plate resistance r_p of the tube at the corresponding plate voltage. The result is solved for ϵ, giving

$$\epsilon = \tfrac{3}{2} i r_p - e_b \qquad \text{volts} \tag{32}$$

where i and r_p are expressed in amperes and ohms, respectively.

The importance of the correction voltage ϵ in connection with tubes containing grids will be shown in the discussion of the transconductance of triodes.

24. Triodes. In triodes the space-charge-limited cathode current is controlled in a manner analogous to that in the diode. In a triode, or a tube with more than one grid, the voltage of each electrode contributes to a composite controlling voltage of an equivalent diode anode, the dimensions of which are influenced by the dimensions of all of the actual electrodes except the cathode. The expression for cathode current is

$$i = G(e')^{3/2} \tag{33}$$

where G is the perveance of a diode having the same cathode and an anode of the same dimensions as the equivalent diode anode. The voltage e' is the composite controlling voltage.

For a triode having a high amplification factor, the equivalent diode anode may be taken as coinciding with the grid. Where such an assumption is not warranted, the following relations for locating the equivalent diode anode are generally of sufficient accuracy for design purposes.

1. *Plane electrodes:*

$$x' = (\mu x_1 + x_2)/(\mu + 1) \tag{34}$$

where x' = distance between cathode and equivalent diode-anode
x_1 = distance between cathode and grid plane measured to center of grid wire
x_2 = distance between grid plane and plate
μ = amplification factor[1]

Where the cathode is a plane parallel to the grid, the perveance is that given by Eq. (20) when x' is substituted for x. Where the grid and plate are parallel planes and the cathode is a wire, x' is taken as the distance h between the cathode axis and a parallel plane diode-anode. The reader is reminded that in computing the current, Eq. (21) may be used by substituting $2r$ for h as was discussed in connection with Eq. (26).

As a singular case, the equivalent diode-anode of a triode in which all three electrodes are parallel planes is located almost precisely in the plane of the grid if the composite controlling electrode voltage is expressed in the form

$$e' = \frac{\mu e_c + e_b}{\mu + 1 + \frac{4}{3}(x_2/x_1)} + \epsilon \tag{35}$$

2. *Cylindrical electrodes:* the radius r' of the equivalent diode-anode is given with reasonable accuracy by

$$r' = \log^{-1} [(\mu \log \rho_p + \log \rho_g)/(\mu + 1)] \tag{36}$$

where ρ_g and ρ_p are, respectively, the radii of the grid (to center of grid wire) and of the plate. The value of r' given by Eq. (36) is used directly in Eq. (21) for perveance.

3. Where the grids are *oval* and the plate is either oval or circular in cross section, certain approximation methods are useful. The distance between the equivalent diode-anode and cathode axis is taken in the vertical plane where the grid-cathode spacing is minimum, and computed as for plane electrodes.

Since the average spacing is greater than the minimum, an experimentally determined correction factor must be employed; usually it is about 0.7. A somewhat more accurate method is to divide the cross section of the tube into sectors and compute each sector separately. The current is then the sum of the currents in each sector, the mutual conductance is the sum of the mutual conductances of each sector, and the amplification factor is given by solving Eq. (47).

The composite controlling electrode voltage e' for triodes, except where Eq. (35) applies, is given by the relation

$$e' = \frac{\mu e_g + e_p}{\mu + 1} \tag{37}$$

where e_g and e_p are the grid and plate voltages, respectively. Where e' is of the order of 10 volts or more, it is generally satisfactory to take the electrode voltages applied to the tube terminals, *i.e.*, $e_g = e_c$, $e_p = e_b$, where e_c and e_b are, respectively, the applied grid and plate voltages.

Where e' is low, as in receiving tubes having a high amplification factor, e_g and e_p of Eq. (37) must be taken to mean that correction voltages are included. Although the contact potentials of grid and plate are usually not identical, it is convenient to disregard this fact, the error in so doing being negligibly small. In this case the composite controlling voltage may be written in terms of the applied voltage and a single correction voltage in the form

$$e' = \frac{\mu e_c + e_b}{\mu + 1} + \epsilon \tag{38}$$

25. Evaluation of Correction Factor ϵ. As will be shown later, the value of ϵ is a factor in the determination of the characteristics of the tube: hence means are required

[1] For design purposes, the amplification factor used is the mu factor (μ_{pgk}), meaning the magnitude of the quotient of plate voltage change by grid voltage change for constant cathode current.

for its evaluation. For purposes of design, it is computed from static and dynamic readings. For purposes of quality control an approximate value is obtained by a simple measurement. Both methods are described below.

1. For design applications, the value of ϵ is readily found from a set of bridge readings made under the conditions for which ϵ is to be determined. In a triode, the cathode current is the same as in a diode when the equivalent diode-controlling voltage e' as given by Eq. (37) is substituted for e_p in Eq. (19). The expression becomes

$$i = G\left[\frac{\mu e_c + e_b}{\mu + 1}\right]^{3/2} \tag{39}$$

Now if the grid is negatively biased, the current i is the plate current i_b and the transconductance g_m is the rate of change in plate current with change in grid voltage, plate voltage held constant. Assuming the perveance G in Eq. (39) to be constant with change in grid voltage, the derivative di/de_c is taken and G eliminated. This results in the useful relation

$$\epsilon = \frac{3i\mu}{2g_m(\mu + 1)} - \frac{\mu e_c + e_b}{\mu + 1} \quad \text{volts} \tag{40}$$

Thus the measured values of plate current, amplification factor, and transconductance at the voltage conditions of measurement may be set into Eq. (40) to obtain ϵ. The value so obtained is in most cases slightly too high because G was taken to be constant, but the error usually cancels out when such values of ϵ are put back into a design equation where the same assumption is allowed. Attention is called to the fact that, if the current i is expressed in milliamperes, g_m is to be expressed in millimhos.

Fig. 9. Circuit for rapid check of grid internal correction voltage. A microammeter may be placed in the grid circuit for checking both positive and negative values by the slower method.

In most cases where ϵ is of importance, *i.e.*, in high-gain receiving tubes, its value is likely to be between +0.3 and +1.0 volt. Where a value in excess of 1 volt is noted, the tube may be poorly aged. In tubes having gold-plated control grids, as in some computer types, the correction voltage is sometimes negative. For amplifier tubes used in "hybrid" autosets, ϵ is of special importance because the available battery voltage is so low. Further comments on the correction voltage will be found in the paragraphs dealing with transconductance and converter tubes.

2. While Eq. (40) gives the value of ϵ with accuracy necessary for purposes of design or redesign, test results satisfactory for quality-control purposes may be obtained using the circuit shown in Fig. 9. The resistor R ordinarily has a value of 10 megohms, and the plate supply is adjusted to a value that results in a plate current of approximately rated value for the average tube when the grid return is through resistor R (switch in position b). The cathode is heated at rated operating voltage, and the plate current is carefully noted. The switch is then thrown to position a and the plate current returned to its original value by adjustment of the potentiometer. The voltmeter then reads the internal correction voltage ϵ, sometimes incorrectly referred to as the "contact potential." It should be noted that for directly heated cathodes the test should be made with direct current only and with the grid return made to the negative end of the filament.

The correction voltage is usually positive, but tubes are occasionally found in

which the correction voltage is of negative polarity. Where knowledge of the magnitude of such a voltage is required, a microammeter may be placed in the grid circuit and the voltmeter reversed so that a positive voltage is applied to the grid. The voltage at which grid current just begins to flow is then noted. In either case, the correction voltage is minus the voltage being applied by the potentiometer.

26. Mu Factors. In general, the mu factor μ_{jlm} of an electron tube is the magnitude of the quotient of the alternating voltage of electrode j by the alternating voltage of electrode l under the condition that the current to electrode m remain constant. In amplifier circuits, the amplification factor μ_{pgp} is of most general interest, although, as has been noted, μ_{pgk} is used in tube design, since it is the total cathode current that is directly related to the composite controlling voltage.

In certain cases of restricted interest the mu factor μ_{gpg} of a triode with positive grid and negative plate is taken, the value of the factor being less than unity. It should be pointed out that in such operation the perveance G is considerably reduced by electrons that pass twice through the grid, the reason being that an electron moving in the grid-to-cathode direction has the same effect upon the space charge as an electron moving in the usual direction.

In the paragraphs immediately following, it is assumed that $\mu = \mu_{pgk}$ or, of course, μ can be μ_{pgp} if it is assumed that the grid draws no current from the electron stream and that the frequency is low.

Several formulas have been published for μ in terms of the dimensions of the tube. Those of Vogdes and Elder are commonly used and are given below. The distance between grid and cathode and the shape of the cathode are not factors, although they do have some influence when the spacing or the composite controlling electrode voltage is low.

For parallel-plane electrodes,

$$\mu = k_1 x_2 - k_2 \qquad (41)$$

For coaxial cylinders,

$$\mu = k_1 \rho_g \ln (\rho_p/\rho_g) - k_2 \qquad (42)$$

where x_2 = grid-plate spacing for parallel electrodes
ρ_g = radius of grid
ρ_p = radius of plate
$k_1 = 2\pi n/\ln \coth 2\pi nr$
n = grid turns per unit length
r = radius of grid wire
$k_2 = (\ln \cosh 2\pi nr)/(\ln \coth 2\pi nr)$

It is seen that, for plane electrodes, μ varies directly as the grid-plate spacing. For cylindrical electrodes, differentiation of μ with respect to ρ_g shows that μ is maximum when $\rho_g = \rho_p/2.718$, which means that, for this ratio, accidental variations in ρ_g or ρ_p result in minimum variation in μ.

Values of k_2 are seen not to vary with x_2, ρ_g, or ρ_p and are small in magnitude; hence they are frequently ignored in adjusting the dimensions of a tube to achieve an alteration in μ. Figure 10 provides a convenient means for the evaluation of k_1 in terms of wire size in inches and winding in turns per inch.

It is often necessary to design triodes in such a manner that the distance between the grid and anode is varied circumferentially, *e.g.*, where the grid is oval and the anode either cylindrical or rectangular in cross section. It is evident that the density of the electron stream is likely to be greatest where the grid-plate spacing is smallest, and so that part of the structure will largely determine the amplification factor. For such a tube, a preliminary design is tested in which the minimum distance is taken as x' in Eq. (34); final adjustment will be based upon an effective value of x' determined by measurement of μ for the preliminary samples.

27. Mu-slop and Cutoff. When the anode current is reduced by either decrease in anode voltage or increase in negative grid voltage, μ is observed to decrease. This decrease is generally not great where the effective controlling electrode voltage is of

the order of several volts, but as the current is decreased to the order of some microamperes the decrease in μ becomes pronounced. This effect, known as mu-slop, is due to a number of causes. Slight variations in dimensions, such as the misplacement of a grid wire or incomplete coverage of the plate by the grid, may permit the passage of current through one part of the tube when it has been cut off at all other parts. The effect is that the μ of the tube is, under this condition, the same as though the entire tube were made with a more open-spaced grid.

Mu-slop plays an important part in the determination of cutoff, although the extent is not always easy to determine. Reference to Eq. (39) indicates that the current is zero when the bracketed term is zero. However, as the term approaches zero, μ decreases and ϵ increases,[1] neither change being readily predictable. Mu-slop is most pronounced in receiving tubes having indirectly heated cathodes while the change in ϵ is greatest for tubes having high emission. Furthermore, the mu-slop at a given value of plate current increases as the plate voltage is increased. In view of these uncertainties, it is not possible to predict the precise value of grid voltage at which the current shall be some arbitrary low value. However, the following empirical relation is used in setting up a quality test for cutoff voltage for a current maximum of 10 μa:

$$e_{co} \cong -1.35 e_b/\mu \tag{43}$$

where e_{co} is the grid voltage applied and e_b and μ are the rated values. For pentodes, e_b is replaced by e_{c2} and μ by the triode mu, μ_1.

28. Transconductance and Plate Resistance. Differentiation of Eq. (39) also provides the relation which gives the grid-plate transconductance or mutual conductance g_m in terms of amplification factor, plate current, and electrode voltages. Equation (39) is to be taken with e_c sufficiently negative so that all of the current goes to the plate. The result may be expressed in the form

$$g_m = \left(\frac{di_p}{de_c}\right)_{e_c} = \frac{3}{2} G \left(\frac{\mu}{\mu + 1}\right) \left(\frac{\mu e_c + e_b}{\mu + 1} + \epsilon\right)^{1/2} \tag{44}$$

where i_p is the plate current, the grid current being zero. Equation (44), with elimination of G by means of Eq. (39), gives the useful relation

$$g_m = \frac{3}{2} \frac{\mu i_p}{\mu e_c + e_b + \epsilon(\mu + 1)} \tag{45}$$

from which Eq. (40) was taken by rearrangement of the terms. Note that if i_p is expressed in milliamperes, g_m is in millimhos.

In the discussion of diodes it was indicated that the perveance formulas do not hold when the anode-cathode distance becomes very small. This is also true in tubes having grids, so that Eqs. (44) and (45) also fall down under such conditions. The maximum theoretical value of g_m is of the order of 10^4 μmhos per ma.

Expressions for the plate resistance r_p may be derived just as Eqs. (44) and (45) were obtained by taking $1/r_p = (di_p/de_b)$, the result being merely a demonstration of the well-known rule that

$$r_p = \mu/g_m \tag{46}$$

A relation that is very useful in tube design is obtained from Eq. (45) for determining the amplification factor necessary when the values of g_m, i_b, e_c, and e_b are specified, which constitutes one of the first steps necessary in the design of any tube having one or more grids. The equation is

$$\mu = \frac{2g_m(e_b + \epsilon)}{3i_b - 2g_m(e_c + \epsilon)} \tag{47}$$

[1] The potential minimum, an important part of ϵ, is seen in Eq. (30) to vary as i varies; as i decreases, V_{min} increases negatively while the contact potential difference, $-(\phi_a - \phi_k)$, presumably remains constant; thus ϵ changes in the positive direction.

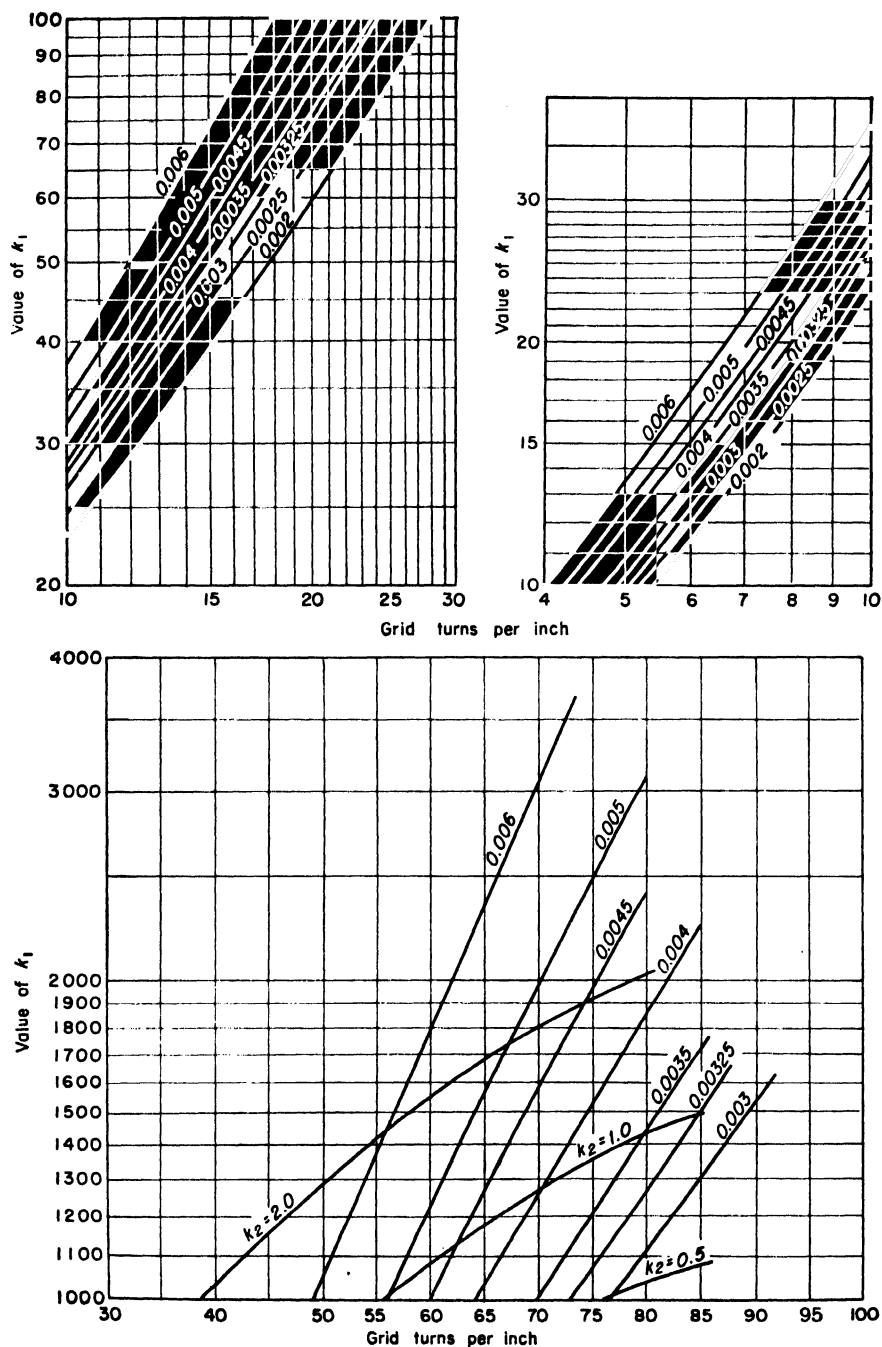



FIG. 10. Evaluation of k_1 in terms of wire size in inches and winding in grid turns per inch.

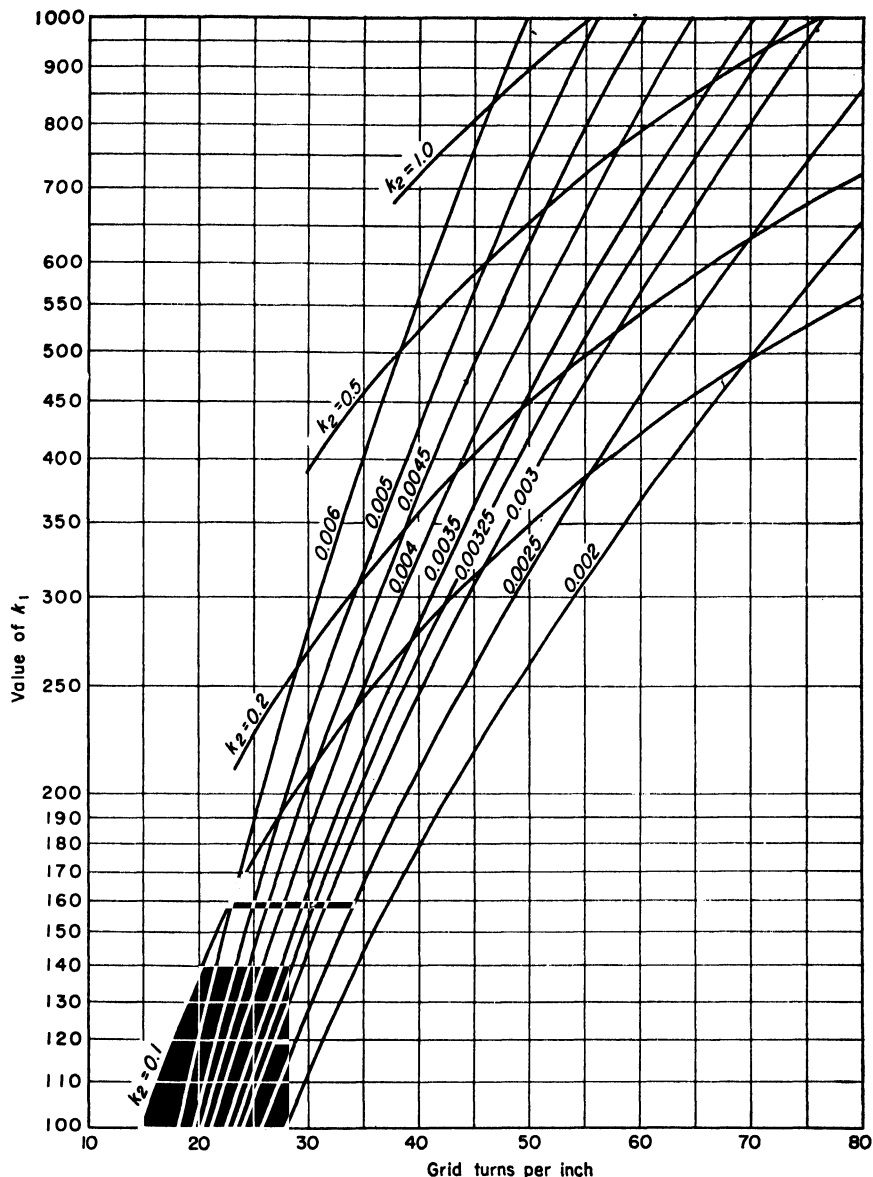



FIG. 10. (*Continued*)

Inspection of Eq. (47) shows that, for required values of g_m, i_b, e_c, and e_b, a high value of ϵ requires a higher value of μ than would be required when ϵ is low. It is, therefore, important to know at the start, especially in a high-mu triode or a high-gain, sharp cutoff pentode, how low a value of ϵ may be expected from the tube when well aged or what its value will be after some hours of operation.

Control of ϵ. As previously noted, the expected value of ϵ cannot be predicted with accuracy, but reasonable estimates of the probable value can be based on similar tubes. Variations from average for a type, whether the variation is among individual tubes or changes during life, are important considerations in the control of quality of high-gain tubes.

Equation (40) assumes full space-charge-limited current, but when the current of a tube is partially temperature-limited, the quotient of plate current by transconductance is high; hence such a tube may be expected to show a very high value of ϵ when so computed. For this reason, tubes which apparently have an internal correction voltage much above 1 volt should be checked for emission.

Apart from the effect of poor or unstable emission, the principal cause of variation in ϵ is usually the variation in contact potential of the control grid. In tubes having coated cathodes this variation may be as much as 1 volt; in tungsten-filament tubes it is usually small. Where a coated cathode is used, the grids appear to become contaminated by the cathode during life so that the actual contact potential difference approaches a very low value; hence ϵ approaches the value of $-V_{\min}$. In most indirectly heated tubes, ϵ will become stable after the tube has been in use for 50 or 100 hr. In high-mu triodes and in high-gain receiving-type pentodes with coated cathodes the grid contact potential may be partially stabilized by taking tubes initially at room temperature and raising the heater voltage fairly rapidly to about twice the rated value and holding the condition for about 30 sec, then turning off the current. No other voltages are applied and the cathode must not be heated longer than indicated, since otherwise the grid becomes heated and the condition is reversed. The process may be repeated a number of times, usually with some continued change in ϵ, provided sufficient time is allowed between treatments for the grid to cool off. The treatment is the same regardless of whether ϵ is originally higher or lower than the stable value finally obtained.

In tubes having coated cathodes made for industrial purposes, where constant characteristics over a period of some thousands of hours is a primary requisite, special means are employed for maintaining the contact potential at its initial value. The cathodes are frequently designed so that the temperature is somewhat lower than that of similar radio receiving tubes. This reduces the rate at which coating is volatilized; hence the maintenance of the coated diameter and activity of the cathode is improved and the rate of grid contamination is reduced.

29. Division of Current in Positive-grid Triode. The manner in which the total cathode current of a positively biased tube divides between grid and anode cannot be specified with any high degree of accuracy. There are, however, formulas that yield fair results in the absence of grid primary or secondary emission and are based on the following concepts:

1. When the grid is at its "natural" potential, the ratio of grid current to plate current is the quotient of grid wire diameter by the distance between grid wires.

2. The ratio of plate current to grid current varies approximately as the square root of the ratio of plate voltage to grid voltage.

The term *natural potential* means the voltage E_γ which, applied to the grid when the plate is at the arbitrary voltage E_π, results in the same cathode current as would have been obtained with the grid physically removed. If the *cathode, grid, and plate are all planes,*

$$E_\gamma = E_\pi \left(\frac{x_1}{x_1 + x_2} \right)^{4/3} \tag{48}$$

where x_1 and x_2 are, respectively, the cathode-grid and grid-plate spacing.

When cathode, grid, and plate are all coaxial cylinders, the natural potential of the grid is

$$E_\gamma = \left(\frac{\rho_g \beta_g{}^2}{\rho_p \beta_p{}^2} \right)^{2/3} E_\pi \tag{49}$$

where ρ_g and ρ_p are, respectively, the grid and plate radii and β_g^2 and β_p^2 are obtained from Fig. 7 using ρ_g/r_0 and ρ_p/r_0, respectively, r_0 being the cathode radius.

For these two cases, one may then take as an approximation, in the absence of primary or secondary emission from the grid, the empirical relation:

$$\frac{i_b}{i_c} \simeq \left(\frac{s - d}{d}\right) \left(\frac{e_b E_\gamma}{e_c E_\pi}\right)^{1/2} \tag{50}$$

where i_b and i_c = plate and grid currents
e_b and e_c = plate and grid voltages
s = spacing between grid wires on centers
d = grid wire diameter

For the case in which the cathode is a wire or ribbon, grid and plate are planes, Spangenberg gives the relation:

$$\frac{i_b}{i_c} = \frac{L + M(E_b/E_c)}{P - Q(E_b/E_c)} \tag{51}$$

where $L = \pi s p \mu - s\mu(p + f)D - 2\pi p r_g \mu$
$M = \pi s p - 2\pi p r_g + sfD\mu$
$P = s(p + f)\mu + 2\pi p r_g \mu$
$Q = sfD\mu - 2\pi p r_g$
s = spacing between grid wires
r = grid wire radius
p = grid-plate spacing
f = filament-grid spacing
μ = amplification factor of tube
$D = r_g/2f \ln (4ef/r_g)$, where $e = 2.718$

30. Tetrodes. Tubes having two grids between cathode and anode are the more common types of tetrodes. The grid nearest the cathode, designated g_1, is ordinarily operated with negative bias and the second grid, g_2, is used to provide increased gain or power output and reduced control grid-plate capacitance. Such tubes suffer by comparison with pentodes in their performance, owing largely to the fact that secondary emission reduces the plate resistance r_p, therefore the gain; the secondary emission varies with life, causing a shift in tube characteristics. In addition, the feedback capacitance is lower in a pentode, thus permitting full advantage of the higher r_p.

The perveance G and internal correction voltage ϵ of a tetrode may be computed or measured provided the conditions are such that electrons are not made to reverse their directions and affect the space charge in the region of the cathode. Under ordinary conditions a small change in plate voltage produces a relatively insignificant change in cathode current. The change in plate current with small change in plate voltage is usually found to be accompanied by a change in screen-grid current which is opposite in direction and almost equal in magnitude, while the change in cathode current is very small indeed. Hence the perveance is practically the same as if the screen grid were replaced by an anode, the potential of this anode being that of the tetrode screen grid. Under these conditions,

$$i = i_b + i_{c2} = G\left(\frac{\mu_1 e_{c1} + e_{c2}}{\mu_1 + 1} + \epsilon\right)^{3/2} \tag{52}$$

where μ_1 = triode amplification factor
i_b = plate current of tetrode
i_{c2} = screen-grid current
e_{c1} = control-grid voltage
e_{c2} = screen-grid voltage

The value of G is then related to dimensions as in the case of a triode. For determining the perveance and the internal correction voltage of an experimental tube from bridge

measurements, the screen grid and plate are connected at the socket and the normal screen-grid voltage is applied. The readings are then substituted in Eq. (52), $(i_b + i_{c2})$ now being the total triode plate current. Similarly, ϵ is obtained when the same data are placed in Eq. (40), the value of g_m now being that obtained for the triode connection and e_b being the voltage applied to the screen grid and plate tied together.

The transconductance of the tetrode is related to that of the triode as in Eq. (53) below, assuming that there will be zero change in the proportional division of current between screen grid and plate when a signal is applied to the control grid, and that the cathode current is the same for either triode or tetrode connection.

$$g_{m(\text{tetrode})} = (i_b/i_k) = g_{m(\text{triode})} \tag{53}$$

where i_b = tetrode plate current
i_k = total cathode current $(i_b + i_{c2})$

The relation $g_m = \mu/r_p$ holds for a tetrode as for any other tube. However, the values of μ and r_p are not readily predictable from the dimensions in view of the fact that the ratio of screen current to plate current changes when a signal is applied to the plate. This change acts to reduce r_p and μ. If the change were zero, the tetrode amplification factor would be approximately equal to the sum and product of (1) the mu of the tube as a triode with the screen grid omitted and (2) the mu of the tube as a triode if the screen grid were the only grid in it. To keep the amplification factor as high as possible, the distance between screen grid and plate is made relatively large and the screen grid may be treated to hold the secondary emission low, as by spraying it with graphite. In transmitting tubes the wires of the screen grid are sometimes placed directly behind the wires of the control grid, thus achieving an increase in plate resistance and reduction in screen-grid dissipation.

31. Pentodes. A pentode is an electron tube having five electrodes: usually a cathode, an anode, and three grids. This discussion will be limited to the case where the electrons from the cathode pass successively through all three grids to the anode. In practically all cases the pentode may be considered as a tetrode in which a third grid is added between screen grid and plate for the purpose of suppressing secondary emission; hence the third grid is often called a "suppressor" grid. The principal results of suppressing secondary electrons are an increase in the amplification factor (hence an increase in gain), increase in power output, and decrease in distortion for large-signal tubes.

The manner in which the third grid suppresses the flow of secondary electrons is readily understood. With the suppressor at zero potential while both screen grid and plate are positive, the field toward the suppressor is retarding for electrons on both sides. Since the average initial velocity of most of the secondary electrons is equivalent to only a few electron volts, the energy is generally too low to allow them to pass through the suppressor. To a certain extent, the control of the secondary electrons is analogous to cutoff in a triode. For example, if μ_{3p} is the suppressor grid-plate mu, one could expect the flow of secondary electrons from the screen grid to the plate to be blocked when $(e_b - e_{c2}) + \mu_{3p}(-e_{c2}) = 0$. In the practical case μ_{3p} is made considerably greater, a minimum figure being

$$\mu_{3p} = 2(e_b - e_{c2})/e_{c2} \tag{54}$$

in which e_b is to be taken as the highest value of plate voltage likely to be encountered under full-signal operation. For ordinary receiving-type pentodes, μ_{3p} is about 15 to 20 when measured at $e_{c2} = 100$, $e_b = 250$, $e_c = 0$.

Figure 11 shows a typical curve of the electrode currents in a pentode plotted against suppressor voltage. Although the effect of the suppressor-grid voltage on the anode current is usually of no importance in ordinary amplifier operation, it is of interest in certain circuits such as multivibrators and square-wave generators. It is noted that the i_b-e_{c3} curve has a flection point at which the suppressor grid-plate transconductance is maximum. This corresponds approximately to the conditions under which half of

the electrons moving toward the plate are turned back by the suppressor grid. Under conditions approaching perfect symmetry, including alignment of the control and screen grids and a high value of μ_{3p}, the i_b-e_{c3} curve becomes very steep at the inflection point—which will occur with a very small negative bias on the suppressor grid—and the suppressor grid-plate transconductance approaches a theoretical maximum value which is of the order of 10,000 μmhos per ma plate current.

In Fig. 11 the cathode current has been plotted against suppressor-grid voltage. It is seen that, where the plate current increases rapidly with decrease in suppressor

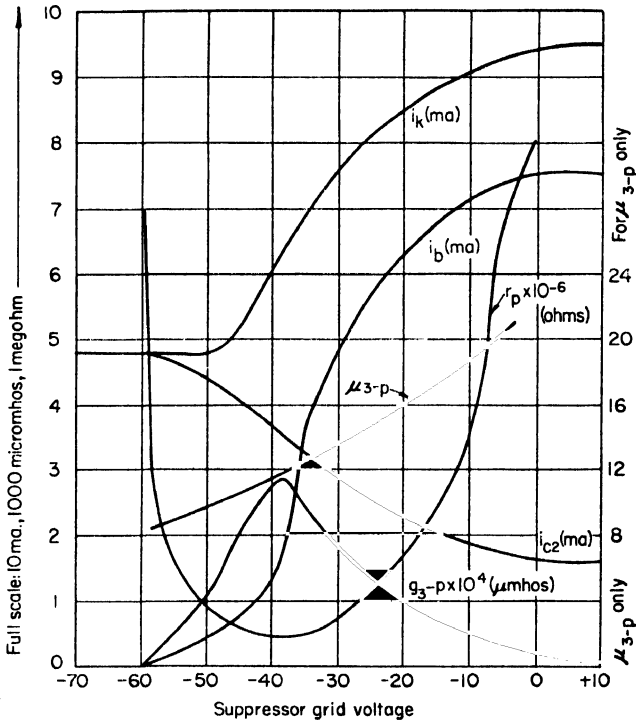



FIG. 11. Variation in i_b, i_{c2}, i_k, μ_{3-p}, and r_p with variation of suppressor-grid voltage.

bias, the total cathode current also rises. This may be interpreted as showing the manner in which the cathode current is limited by electrons returned toward it from beyond the control grid, or, this being granted, as showing that some of the electrons repelled by a bias on the suppressor grid may again pass through both screen and control grids. Figure 11 was taken from a tube having substantially cylindrical electrodes. Had the filament been a wire and the other electrodes flat, the effect of the suppressor-grid bias on the cathode current would have been less noticeable.

The perveance of a pentode is taken in the same manner as in the case of a tetrode, the suppressor—where it has an external connection—being connected to the cathode. The equivalent diode-anode is then just beyond the control grid as in the triode, so that the data given in the discussion of diodes and triodes will apply here.

Similarly, the equivalent controlling electrode voltage may be taken by considering the screen grid to be equivalent to a triode plate; hence,

$$e' \simeq \frac{\mu_{12}e_{c1} + e_{c2}}{\mu_{12} + 1} + \epsilon \qquad (55)$$

32. Pentode Design. While it is beyond the scope of a handbook to outline in detail the design of a pentode, mention will be made of a few points which may aid in an understanding of the more difficult steps. It is assumed that the objective is a tube having specified values of i_b, i_{c2}, and g_m at specified values of e_{c1}, e_{c2}, and e_b.

1. The first step is to decide on a value of the internal correction voltage. If the tube is to resemble an existing tube to the extent that the values of ϵ could be expected to be alike, the existing tubes can be read as triodes and ϵ computed from Eq. (40). In the absence of such data, ϵ can be taken arbitrarily as, say, 0.7 volt in the preliminary test.

2. The required triode mu is computed from Eq. (47), using e_{c2} in place of e_b and e_{c1} in place of e_c. (Note that the screen current does not enter the relation, since the pentode g_m and i_b are the same fractions of the triode g_m and i_b, respectively.)

3. Assuming that the general configuration of the electrodes and the dimensions of the cathode have been picked, the dimensions of the composite anode are chosen to give the proper perveance as described in the paragraphs on diodes. The value of the perveance required will be

$$G = \frac{i_b + i_{c2}}{\left(\dfrac{\mu_{12}e_{c1} + e_{c2}}{\mu_{12} + 1} + \epsilon\right)^{3/2}} \tag{56}$$

where μ_{12} is the required triode amplification factor as computed by the method outlined in the preceding paragraph.

4. The surface dimensions of the control grid are first taken as being those of the composite anode, and the surface dimensions of the screen grid then chosen to have reasonable clearance and shape. The exact surface dimensions of the control grid may then be determined as for a triode. See Eqs. (34), (36), (41), (42), and accompanying text. When these results are at hand, good judgment will frequently indicate that a change may be necessary in some of the parts arbitrarily chosen; *e.g.*, a larger cathode might better be used, or a screen grid of smaller diameter might have some advantage. In the former case, a new composite anode must be computed; in the latter case, a new control grid will be computed for the smaller screen grid.

5. No attempt is made in the first design to produce a specified screen-grid current. The first batch of tubes must, in any case, be regarded as means to provide a basis on which to make corrections, and such corrections may include the correction for screen-grid current. In a shadow-grid (grid turns aligned) type of tube, the screen-grid current changes in the same direction as the diameter of the wire with which it is wound and, other factors being the same, increases as the distance between g_1 and g_2 is increased.

6. The dimensions of the plate are not especially critical so far as the design is concerned, but this is not to be taken as meaning that plates for a given design can be allowed to vary over wide limits. The plate is to be shorter than the grids to prevent the field at the ends from extending around the grids and influencing the electron current or resulting in unnecessarily high control grid–plate capacitance.

7. In view of Eq. (54) the mu of suppressor to plate must be adequate. On the other hand, μ_{3p} must not be too high for the reason that the plate resistance of the tube may be too low, especially when operating under the condition that $e_{c2} \simeq e_b$. The reason for this effect is discovered when it is considered that electrons, in passing close to the screen-grid wires, may have their paths bent in such a way that, while their total momentum is unchanged, the component in the direction of the retarding field of the suppressor is reduced to the point where a slight difference in plate voltage will have a major effect upon the probability that such an electron shall reach the plate. The writer knows of no formula that accurately predicts the dimensions of the ideal suppressor grid. When the suppressor grid has an external connection, some check may be obtained by measuring the plate resistance r_p of the tube with the potential of the suppressor changed a few volts above and below zero; if maximum r_p is

obtained with the suppressor negative, it is generally an indication that it would be advantageous to wind it with a tighter pitch; if r_p is maximum with the suppressor positive, the pitch may be opened. In any event, the final pitch selected is the result of comparative tests.

33. Shielding. In a pentode intended for r-f operation, shielding is used primarily for the reduction of control grid–plate capacitance. The design of shielding elements in a tube cannot be said to be based on formula. The evolution of shielding since the appearance of the first shielded tetrodes about 1925 might be considered as an example of intelligent cut-and-try engineering, particularly as exemplified in some of the high-perveance, single-ended pentode receiving tubes. In constructing experimental tubes the engineer who does not have available the advice of persons having experience in this phase of tube design will probably obtain best results if control grid and plate leads are taken out at opposite ends of the tube and the design features of commercial tubes, so far as shielding is concerned, are carefully copied. In such cases, the shields of discarded tubes are often found to be worth salvaging for experimentation.

34. Remote-cutoff Tubes. Ordinarily, remote-cutoff tubes are designed to fit a preferred curve of control-grid voltage plotted against either transconductance g_m or conversion transconductance g_c.

The control grid may be wound with a continuously variable pitch, the most open part being in the center. It is more practicable, however, in pentode amplifiers to use only three pitches—the most open being in the middle, then one or a few turns of an intermediate pitch each side of center, and a constant, close pitch over the balance of the length of the grid.

So far as the pentode is concerned, an approach to the problem of design is made by considering that the tube has three parts: one part having a high value of triode mu, one part having an intermediate value, and a third part having a low value. The length of each part of the tube is the same as that of the corresponding parts of the control grid. In the five paragraphs immediately following, the problem is considerably oversimplified, although it does show in outline a means for setting up a preliminary test. Final design is achieved, as with other tubes, by adjustment of factors that may be computed by measurement of the preliminary samples.

For the pentode amplifier, suppose three points on the g_m-e_{c1} curve are fixed at the bogey values that the design is expected to achieve. Let $g_{m(1)}$ be the mutual conductance at the low-bias end of the curve, the control-grid voltage being $e_{c(1)}$; at an intermediate point the values are $g_{m(2)}$ and $e_{c(2)}$ while at a point near cutoff they are $g_{m(3)}$ and $e_{c(3)}$. Correspondingly, there will be three values of plate current: $i_{(1)}$, $i_{(2)}$, and $i_{(3)}$.

At grid bias $e_{c(3)}$, the tube will be cut off all except at the one turn in the center. Here the triode mu, $g_{m(3)}$ and $i_{(3)}$, and the voltages will be related as in an ordinary triode or pentode, and the pitch of the control grid, n turns per unit length, is computed. The length of this part of the tube is $1/n$, and its current and transconductance for this section may then be computed to a reasonable degree of accuracy at any other value of grid bias.

At the intermediate point the current and transconductance of the low-mu section are computed and subtracted from the required values. The difference between this and $i_{(2)}$ and $g_{m(2)}$ is then made up by computing the length and mu of the intermediate section to give a total current and transconductance of the proper value for this point.

Similarly, at the low-bias end, the current and transconductance of the low and intermediate sections are computed and subtracted from the required values $i_{(1)}$ and $g_{m(1)}$. The difference is then taken and the high-mu section computed to bring the totals to the required values.

Such a procedure will often result, on first trial, in a computed cathode length which is not satisfactory as a practical design. The perveance per unit length may then be raised or lowered, as by altering the surface dimensions of the control grid, and a new set of computations made.

35. Beam-power Tubes. Also classed as pentodes are the so-called beam-power output tubes employing a cathode, anode, two grids, and a pair of beam-confining

plates, the latter taking the place of the suppressor grid in the high-gain pentode. The word *beam* is used to indicate that the wires of g_1 are interposed between the cathode and the wires of g_2 so that the electron stream is in the form of beams passing toward the anode. Figure 12 shows in cross section the electrode arrangement of a typical tube of this variety.

In a beam-power output tube the beam-confining plates are ordinarily connected internally to the cathode. While the relations between voltage, current, and mutual conductance and the dimensions of cathode, g_1, and g_2 are as in any other tube, the design from g_2 outward is still largely a matter of cut and try. Achievement of a satisfactory design is not difficult when the problem is understood in a qualitative way.

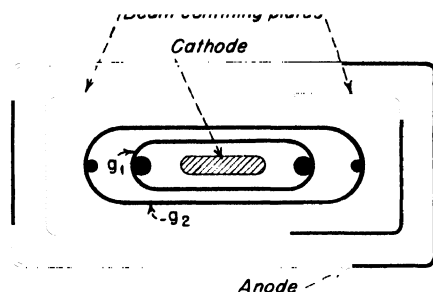



FIG. 12. Cross section of beam-power tube for low-voltage operation. Proper spacing between edges of beam-confining plates is important to performance.

Under full-signal conditions the instantaneous plate voltage of the tube will reach a value much higher than that of the screen grid, and to prevent secondary electrons from g_2 reaching the plate, a retarding field must be provided. Reference to Fig. 12 readily discloses that the retarding field due to the beam-confining plates would be minimum midway between the edges, increasing toward the edges. Therefore, to have sufficient field midway between the edges of the beam plates and not too strong a field toward the edge of the electron beam, the clearance between g_2 and anode should be relatively large and the beam-confining plates placed about halfway between.

When the plate voltage swings to its minimum value, the field between screen grid and plate must also be retarding to prevent the passage of electrons from the plate back to the grid. This adds up to an indicated requirement that the beam-confining plates have a fairly generous clearance on both sides. It will be understood that when a tube is delivering power to a load, the plate current is maximum when the plate voltage is minimum, and that the electrons passing between the edges of the beam-confining plates, because of their negative charge, further reduce the potential in the retarding region to a value that is considerably less than that under a charge-free condition. Now if the edges of the beam-confining plates are too close to each other, the beam is too closely confined and the charge density may become so great that many electrons are turned back toward the screen. The result of such a condition is low power efficiency and high distortion.

In practice, the design of the beam-confining plates is ordinarily settled by making samples with varying spacing between the edges. The tubes are tested for power output and distortion so that, with proper choice of load and a signal whose peak is equal to 95 per cent of the grid-bias voltage, tubes giving best power-output efficiency and lowest distortion are taken as the prototypes. In most cases a power-output efficiency $(100P_0/i_be_b)$ of 50 per cent can be obtained with a total (rms) distortion not greater than 5 or 6 per cent.

In most a-c–d-c radio sets, the screen-grid and plate voltage are taken from the plus end of the power supply, the screen-grid connection being made directly while the plate supply loses some direct voltage due to drop in the output transformer. Under such conditions, with screen-grid voltage somewhat higher than plate voltage, the distortion is apt to be considerably greater than when the plate voltage is equal to or greater than the screen voltage. For this reason the final design of tubes for a-c–d-c sets should be made partially in view of distortion measurements made at, for example, 105 volts at the plate and 110 volts on the screen grid.

36. Converter Tubes. A converter tube is an electron tube which operates as a converter or first detector in a superheterodyne receiver and, distinct from a mixer,

supplies its own local oscillator signal. Some tubes employ two electron streams so that the tube is actually an oscillator and mixer combined; each part may be considered as a triode, pentode, etc., and will not require special mention here.

The pentagrid converter employs a single electron stream, the plate current being varied in magnitude at oscillator frequency by the signal on g_1, and by the r-f signal on g_4. The intelligence contained in the r-f signal is then obtained as a modulation of the i-f signal whose frequency may be the sum of or difference between the oscillator and r-f carrier frequencies.

In the usual pentagrid converter there are a cathode, five grids, and an anode. The second grid frequently comprises a pair of heavy wires disposed between the side rods supporting g_1 and g_3, and is conventionally called a *hairpin* grid. Usually g_3

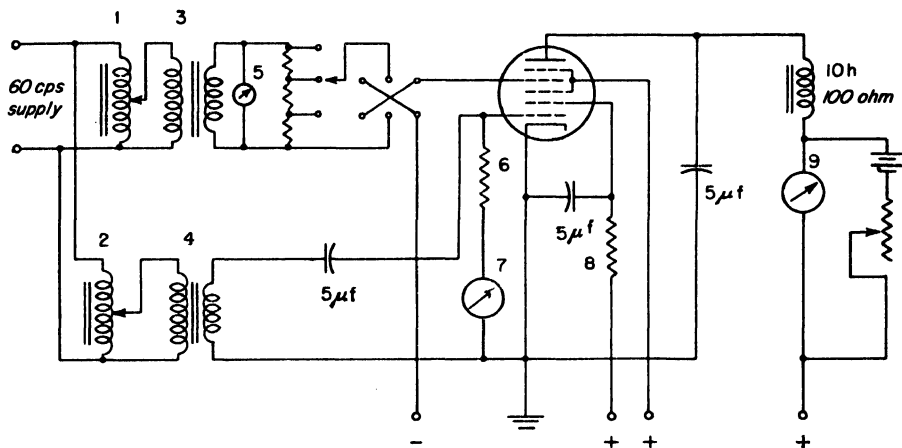



FIG. 13. Circuit for measuring conversion transconductance of pentagrid converter tubes. (1, 2) Variacs. (3) Step-down transformer, 5-volt secondary. (4) Step-down transformer, 25-volt secondary. (5) A-c voltmeter, 5 volts full scale. (6) Resistor of value specified for tube. (7) D-c milliammeter, 1.0 ma full scale. Adjust current by (2) to rated value of tube under test. (8) Dropping resistor; use when specified. (9) Laboratory-type multirange milliammeter.

and g_5 are tied together internally, as both are used as screens, while g_4 is a variable-pitch control grid upon which the r-f signal is introduced. In operation, g_1 and g_2 are used in the same way as the grid and plate of an ordinary triode oscillator with g_1 self-biased.

The conversion transconductance g_c of a converter tube is the quotient of the alternating plate current component of intermediate frequency by the alternating voltage at the signal grid under a standard condition of oscillation. The characteristic is ordinarily measured with a 60-cycle signal frequency, and since this measurement is not commonly given in textbooks, it is described here. The circuit is shown in Fig. 13.

It will be recognized that if the oscillator and signal frequencies are equal, the difference frequency in the plate circuit will be zero, and that the plate current will be greater when the two signals are in phase than when they are applied 180 deg out of phase. Therefore, the difference in direct plate current which may be measured when the phase relation of the two signals is reversed, divided by twice the peak value of alternating voltage applied to g_4, is the conversion transconductance g_c.

Since it is neither convenient nor desirable to have the tube operating as an oscillator to provide both signals for such a measurement, the signal is taken from the 60-cycle supply as indicated in the figure. The magnitude of the signal to g_1 is commonly adjusted for each tube so that the grid current will have the value rated for the oscillation condition at the specified value of the grid resistor 6 in Fig. 13.

A bucking circuit is placed in the plate lead so that the plate current difference Δi_b, obtained when the reversing switch is thrown, may be read accurately; the current in the meter is first adjusted to zero when the signals are out of phase. The conversion transconductance will, with such a test, be given by the relation

$$g_c = \Delta i_b / 2E_{g4}\sqrt{2}$$

where E_{g4} is the rms voltage applied to g_4. With most tubes it is convenient to set E_{g4} at 0.354 volt so that g_c in micromhos will be numerically equal to Δi_b when the latter is read in microamperes. In any event, E_{g4} should be held to a low value, and the wave form of the line voltage should be reasonably free of distortion. The usual precautions with respect to constancy of direct supply voltages and the adequacy of by-passing are to be observed.

A most important design feature of the pentagrid converter is the variable-pitch control grid, g_4. This grid operates at a negative voltage, the magnitude of which is controlled by the a-v-c circuit of the receiver. A curve showing g_c plotted against e_{c4} should, therefore, resemble in a general way the shape of the g_m-e_c curve of the pentode amplifier used in the same equipment. If the winding of g_4 were constant, the g_4-plate transconductance, hence the conversion transconductance, would be sharply peaked at a value of e_{c4} which would depend upon the pitch and upon the dimensions and voltage of g_5. Furthermore, if the grid were made with a number of sections having different pitches, the curve would be full of humps, each hump corresponding to a certain pitch. The reason why this would be so is obvious in the light of what has been said of the suppressor-grid voltage-plate current relations in a pentode. The conditions in the region between g_3 and g_5 of a pentagrid converter are similar in that a negative grid has a positive electrode each side of it, so that control consists in turning back a portion of the electrons that approach it.

In view of the fact that each section produces a peak in g_c at a different value of e_{c4}, if a smooth curve is to be obtained the pitch of g_4 must be varied almost continuously throughout its length. In practice, the most widely spaced turns are in the middle of the grid, the spacing between turns being decreased with a reasonable degree of uniformity from the middle to the ends. The turns at the ends are spaced to result in maximum g_c being obtained at a value of e_{c4} about 1 volt less negative than the bias value designed for minimum signal, *i.e.*, at $e_{c4} = -2.0$ volts in most receiving tubes.

Sometimes control grids g_4 are found having several turns of constant pitch at each end. This practice can be overdone, and it results in a very steep slope in the curve of g_c against e_{c4} in the region of -3 volts with considerable spread in the value under standard test conditions, usually $e_{c4} = -3.0$.

A further result of improper design of g_4 is an abnormally large shift in g_c during the first 50 or 100 hr of operation. In the section on triodes, attention was called to the manner in which the internal correction voltage ϵ influences i_b and g_m, and to the fact that, by proper aging, ϵ can be reduced to a value low enough so that, when the tube is designed for such a value, the deterioration of the characteristics on life is considerably reduced. In a pentagrid converter the slump in g_c during early life is due almost exclusively to change in the contact potential of g_4. Because the grid is so far from the cathode, it does not respond readily to the 30-sec hot shot described for triodes, so that recourse must be had to such help as can be obtained by spraying the grid and by avoiding designs that result in too steep a slope in the g_c-e_{c4} curve in the region of 3 to 5 volts bias.

37. Interelectrode Capacitance. The interelectrode capacitances of electron tubes may be computed to the extent that the end or edge effects may be ignored. The shunt capacitances added by the stem, base, etc., are usually taken by measurement of the complete tube, then of the base and stem with tube electrodes removed. For adjustment of capacitances due to stem and base, a judicious choice of basing arrangement will in many cases be indicated, as well as arrangement and spacing of lead wires in the stem.

The basic formulas, as in the case for perveance, are given first for diodes, then

expanded for tubes having grids. The dielectric constant is unity for vacuum, hence does not appear in the formulas.

For parallel planes,

$$C = 0.0885A/x \qquad \mu\mu f \tag{57}$$

where A = area, sq cm
x = spacing, cm

For coaxial cylinders,

$$C = 0.2416L/\log_{10}(r/r_0) \qquad \mu\mu f \tag{58}$$

where L = length, cm
r and r_0 = anode and cathode radii, respectively

Approximate compensation for edge effects can be made in Eq. (57) by increasing A by one-half the product of periphery and spacing and in Eq. (58) by increasing L by $(r - r_0)$.

For tubes having grids, the cold capacitances are readily expressed in terms of the capacitance between conducting sheets located at the grids and the capacitances between the actual grids and the conducting sheets.

In the *triode*, the conducting surface that replaces the grid, here called the *grid surface*, has capacitance C_1 between it and the cathode, and capacitance between the grid surface and anode is C_2. The grid and grid surface are coupled by C_g, related to C_2 and mu by

$$\mu = C_g/C_2 \tag{59}$$

where μ is taken to be μ_{pgk}.

C_1 and C_2 are calculated from Eq. (57) or (58) depending upon shape, x being the distance between parallel planes, r_0 and r being the inner and outer radii of coaxial cylinders.

The equivalent network gives, for the cold triode, the input capacitance

$$C_{\text{in}} = \frac{\mu(C_1 + C_2)}{\mu + 1 + (C_1/C_2)} \tag{60}$$

the output capacitance

$$C_{\text{out}} = \frac{\mu C_2 + C_1}{\mu + 1 + (C_1/C_2)} \tag{61}$$

and the grid-plate capacitance

$$C_{gp} = \mu C_2/(\mu + 1) \tag{62}$$

For the *tetrode*, C_2 is the capacitance between g_1 and g_2 surfaces, C_3 is the capacitance between the g_2 surface and plate, and the capacitance between the g_2 surface and actual g_2 is $C_{g2} = \mu_2 C_3$, where μ_2 is the amplification factor of the tube when g_1 is omitted. The cold capacitances with g_2 grounded through a by-pass having negligible reactance are

$$C_{\text{in}} = \frac{\mu_1(C_1 + C_x)}{\mu_1 + 1 + C_1/C_x} \tag{63}$$

where $C_x = \dfrac{C_3(\mu_2 + 1)}{1 + C_3(\mu_1 + 1)/C_2}$
μ_1 = triode mu, plate at screen-grid surface
μ_2 = triode mu, g_1 omitted

$$C_{\text{out}} = \frac{\mu_2 C_3 + C_y}{\mu_2 + 1 + (C_y/C_3)} \tag{64}$$

where $C_y = \dfrac{\mu_1 C_2}{\mu_1 + \mu_1(C_2/C_1) + 1}$

$$C_{g1-p} = \frac{\mu_1 C_2}{1 + \mu_1(C_2 + C_x)/C_z} \tag{65}$$

where $C_z = \dfrac{C_2^2 + \mu_2 C_1 C_3}{C_1 + \mu_2 C_3}$

For *pentodes*, the corresponding capacitances are taken by replacing g_3 with the grid surface which is coupled to g_3 through the capacitance $\mu_{4p}C_5$. However, the resultant value of grid-plate capacitance is smaller than the actual end effects even with practicable shielding. For most purposes the input capacitance may be taken as that of the triode with plate at g_2, and the output capacitance as that of a triode with the cathode at g_2, since g_2 will be by-passed to ground.

38. Effect of Space Charge. The presence of electrons in the space between the electrodes increases the capacitance, since the field at the anode is greater than when the interelectrode space contains no electrons. For parallel plates in which one electrode is a cathode and the other a plate to which a space-charge-limited current flows, the capacitance is $\frac{1}{3}$ greater than in the charge-free condition given by Eq. (57). This, however, is not to be taken as meaning that the susceptance increases by the same factor; in fact, the susceptance is lower in the space-charge-limited diode than between the charge-free electrodes except at extremely high frequencies. Thus, for the cold tube, the admittance y for a diode having plane electrodes is

$$y = jb_0 = j\omega C \qquad (66)$$

while for the space-charge-limited diode,

$$y = g_0 + j(\tfrac{3}{5}\omega C) \qquad (67)$$

where $g_0 = 3i/2e_p$, the zero-frequency conductance
ω = angular frequency
C = cold capacitance

In tubes having grids the characteristics are all modified by space charge, especially in the cathode-grid region. For example, Eq. (67) applies to the admittance of the equivalent diode-anode of a tube having plane electrodes, so that an approximation of the altered transadmittance y_{pg} may be written

$$y_{pg} \simeq g_m\,[\cos\,(\tfrac{11}{15}C_1 + C_2) - j\omega\sin\,(\tfrac{11}{15}C_1 + C_2)] \qquad (68)$$

where C_1 is the cold capacitance between cathode and grid plane and C_2 is that between grid plane and plate. It is to be assumed that $e_b \gg e'$, the equivalent diode voltage, so that space-charge effects beyond the grid may be disregarded for a first-order approximation. A more accurate expression for the transadmittance is obtained from the equations that include the effects of transit time, shown in subsequent paragraphs.

39. Transit-time Effects in Nonmicrowave Tubes. At frequencies where the time of flight of an electron between a pair of electrodes is an appreciable part of the time of one cycle of the alternating voltage between the electrodes, the characteristics are no longer described by the simple relations that have been given up to this point.

One visualizes an electron as being emitted from the cathode of an electron tube at a relatively low velocity, then being accelerated by the field of another electrode to a relatively high velocity. Regardless of the path taken or of the velocity it may have had at some previous position in flight, the kinetic energy possessed at any instant is the product of its charge by the *net* change in potential. The velocity in the direction of travel will be

$$v = (2Ve/m)^{1/2} \qquad (69)$$

which, in practical units, is

$$v = 5.93 \times 10^7 V^{1/2} \quad \text{cm/sec} \qquad (70)$$

where V is the potential, e and m being, respectively, the charge and mass of an electron.

When the electron passes from a plane cathode to a plane-parallel anode of a diode and the current is space-charge-limited, transit time will be

$$T = 3x/v \quad \text{sec} \qquad (71)$$

where x is the distance between the electrodes and v is the final velocity given by Eq. (70). Where the current is not space-charge-limited, the transit time is lowered, the value under the condition of zero space charge being two-thirds that given by Eq. (71).

The *transit angle* θ of the region between two electrodes of a tube is the product of the transit time T of an electron through the region and the angular frequency ω of the alternating voltage between the electrodes. For a diode,

$$\theta = \omega T \qquad \text{radians} \tag{72}$$

The *impedance* z of the plane diode when the electron current is space-charge-limited is given by Llewellyn[1] in the form

$$z = (12r_0/\theta_1^4)[2(1 - \cos\theta_1) - \theta_1 \sin\theta_1] \\ -j(12r_0/\theta_1^4)[(\theta_1^3/6) + \theta_1(1 + \cos\theta_1) - 2\sin\theta_1] \tag{73}$$

where $r_0 = de_p/di = \frac{2}{3}(e_p/i)$ at very low frequency.

It is generally more useful to take the reciprocal of Eq. (73) so that the admittance of the gap is had. Thus if Eq. (73) is abbreviated in the form

$$z_{11} = r + jx_1 \tag{74}$$

the reciprocal or admittance will be

$$\begin{aligned} y &= \frac{1}{z} = \frac{r^2}{r^2 + x^2} + j\frac{x^2}{r^2 + x^2} \\ &= g + jb \end{aligned} \tag{75}$$

It is noted that the zero-frequency resistance r_0 is a coefficient in both the resistive and reactive parts of Eq. (74); similarly, the zero-frequency conductance g_0 is part of g and b of Eq. (75). From Table 4 giving values of g/g_0 and b/g_0, the admittance at higher transit angles may be found for values of g_0 and θ_1. Thus, if $\theta_1 = \pi/2$, the admittance $y = 0.9418g_0 + j0.4757g_0$.

Table 4. Values of g/g_0 and b/g_0 Space-charge-limited Diode

Radians	g/g_0	b/g_0	Radians	g/g_0	b/g_0
0	1.000	0	7.07	−0.189	2.981
0.4	0.996	0.120	7.2	−0.211	3.087
1.0	0.977	0.301	7.33	−0.228	3.196
1.4	0.954	0.423	$5\pi/2$	−0.246	3.645
$\pi/2$	0.942	0.676	8.4	−0.168	4.096
1.8	0.923	0.547	8.64	−0.106	4.273
$3\pi/4$	0.867	0.723	9.0	−0.004	4.499
2.8	0.811	0.867	3π	+0.135	4.709
π	0.760	0.982	10.0	+0.264	4.906
3.6	0.681	1.143	10.21	+0.290	4.964
4.0	0.603	1.291	10.47	+0.306	5.033
$3\pi/2$	0.439	1.579	11.0	+0.287	5.177
5.2	0.310	1.802	11.52	+0.215	5.351
5.6	0.197	2.005	11.78	+0.165	5.454
2π	0.000	2.409	12.04	+0.111	5.570
6.8	−0.134	2.773	4π	0.000	5.839

It should be noted that, in Table 4, the value of b/g_0 for θ_1 is zero. This is not, of course, to be taken as meaning that the susceptance can be taken as zero for extremely low transit angles. As the transit angle approaches zero, the admittance approaches

[1] Llewellyn, F. B., and L. C. Peterson, Vacuum Tube Networks, *Proc. IRE*, **32**, 144, March, 1944.

asymptotically the value

$$y_{(\theta\to 0)} = g_0 + (3\theta/10)g_0 = g_0 + j(3/5)\omega C \quad (76)$$

where C is the cold capacitance.

In Table 4 it will be observed that g/g_0 is negative for values of θ_1 between $\theta_1 = 2\pi$ and about 9.0 radians, reaching a value of -0.2464 at $\theta_1 = 5\pi/2$. It is this fact that makes oscillation possible in the Llewellyn diode, for if the circuit is adjusted to have an angular frequency ω and the plate voltage is adjusted by Eq. (71) so that $T_1 = 5\pi/2\omega$, oscillations will be produced, provided the shunt conductance of the external circuit is not greater than $+0.2464g_0$.

40. Transit Angle Effects in Multigrid Tubes. In triodes and in tubes having more than one grid, the transit angles in all regions affect the characteristics of the tube. A full discussion is found in the Llewellyn reference. The manner in which the characteristics of some tubes are affected will be seen in the results given below. These equations may be taken as first-order approximations when they are restricted as follows:

1. The electrodes are parallel planes, edge and lead effects ignored.
2. The cathode current is space-charge-limited, and the virtual cathode is assumed to be at the real cathode.
3. Space charge in the region beyond the control grid is assumed to be negligible.
4. The control grid draws no current from the electron stream.

The transit angles will be taken to have the following values:

Region 1. For the region between the cathode and the first electrode, the transit angle will be that given by Eqs. (71) and (72) for the space-charge-limited diode and may be expressed in the form

$$\theta_1 = \omega T_1 = \omega x_1 \times 10^{-7}/2\sqrt{e'} \quad (77)$$

where x_1 is the cathode-grid spacing and e' is the potential of the equivalent grid plane.

Region 2. For the region between the grid and plate of a triode,

$$\theta_2 = \omega T_2 = 3.37\omega x_2 \times 10^{-8}/(\sqrt{e'} + \sqrt{e_b}) \quad (78)$$

For a tetrode,

$$\theta_2 = \omega T_2 = 3.37\omega x_2 \times 10^{-8}/(\sqrt{e'} + \sqrt{e''}) \quad (79)$$

where x_2 is the spacing between control and screen grids and e'' is the voltage of the equivalent screen-grid plane and may be taken to have the value

$$e'' = (\mu_2 e_{c2} + e_b)/[1 + \mu_2 + (x_3/x_2)] \quad (80)$$

where x_3 is the screen grid-plate spacing and μ_2 is the amplification factor of the tube when g_1 is omitted. The error is usually not great if e'' is taken as e_{c2}, provided the value of μ_2 is high.

Region 3. For the region between g_2 and anode,

$$\theta_3 = \omega T_3 = 3.37\omega x_3 \times 10^{-8}/(\sqrt{e''} + \sqrt{e_b}) \quad (81)$$

Although it would be understood that the equivalent diode controlling voltage e' is to include the internal correction voltage ϵ, the error is not great if it is not included in the screen grid and anode potentials; thus the direct cathode current may be expressed

$$i_k = G\left[\frac{\mu_1 e_g + e''}{\mu_1 + 1 + 4/3(x_2/x_1)}\right]^{3/2} = G(e')^{3/2} \quad (82)$$

where G is the perveance and e_g is the voltage applied to the control grid plus the internal correction voltage.

For triodes,

μ_1 = amplification factor
e'' = plate voltage e_b

For tetrodes,

μ_1 = triode amplification factor
e'' = equivalent voltage at screen grid, taken as e_{c2} where μ_2 is high

The total conductance di_k/de' at zero frequency is taken to be

$$g_0 = \tfrac{3}{2}G(e')^{1/2} = \frac{3}{2} i_k \left[\frac{\mu_1 + 1 + \frac{4}{3}(x_2/x_1)}{\mu_1 e_g + e''} \right] \tag{83}$$

When the transit angles in both regions of the triode and all three regions of the tetrode are just great enough to be significant, the grid-plate transconductance may be taken to have the following values:

For the triode,

$$g_m = \frac{g_0 \mu}{\mu + 1 + \frac{4}{3}\frac{x_2}{x_1}\left(1 + \frac{1}{2}\frac{\theta_2}{\theta_1}\frac{\sqrt{e'} + 2\sqrt{e_b}}{\sqrt{e'} + \sqrt{e_b}}\right)} \tag{84}$$

For the tetrode,

$$g_m = \frac{i_b}{i_k} \frac{g_0 \mu_1}{\mu_1 + 1 + \frac{4}{3}\frac{x_2}{x_1}\left(1 + \frac{1}{2}\frac{\theta_2}{\theta_1}\frac{\sqrt{e'} + 2\sqrt{e''}}{\sqrt{e'} + \sqrt{e''}}\right)} \tag{85}$$

where μ_1 is the triode mu.

The *grid-plate transadmittance* y_{pg}, of which the real part is the transconductance, may be expressed in terms of the l-f values given by Eqs. (84) and (85).

For moderately low frequencies,

$$y_{pg} \simeq g_m(\cos\phi - j\sin\phi) \tag{86}$$

$$\text{where } \phi = \frac{\theta_1\left[\frac{11}{30}(\mu_1 + 1) + \frac{7}{18}\frac{x_2}{x_1}\right] + \theta_2\left[1 + \mu_1 + \frac{4}{3}\frac{x_2}{x_1}\left(1 + \frac{1}{8}\frac{\theta_2}{\theta_1}\frac{3\sqrt{e'} + 5\sqrt{e''}}{\sqrt{e'} + \sqrt{e''}}\right)\right]}{1 + \mu_1 + \frac{4}{3}\frac{x_2}{x_1}\left(1 + \frac{1}{2}\frac{\theta_2}{\theta_1}\frac{\sqrt{e'} + 2\sqrt{e''}}{\sqrt{e'} + \sqrt{e''}}\right)} + \frac{1}{3}\theta_3\left(\frac{\sqrt{e''} + 2\sqrt{e_b}}{\sqrt{e''} + \sqrt{e_b}}\right) \tag{87}$$

For very high values of transit angle,

$$y_{pg} \simeq \frac{i_b}{i_k} g_0 \left(\frac{\mu_1}{1 + \mu_1 + x_2/x_1}\right)\left[\frac{\sqrt{e''} - P\sqrt{e_b}}{j(\theta_2/2)(\sqrt{e''} + \sqrt{e_b})}\right] Q \tag{88}$$

where $P = \cos\theta_3 - j\sin\theta_3$
$Q = \cos(\theta_1 + \theta_2) - j\sin(\theta_1 + \theta_2)$

For the triode, the expressions are obtained directly from Eqs. (87) and (88) by setting $e'' = e_b$ and letting θ_3 approach zero so that the second term of Eq. (87) drops out and the bracketed term of Eq. (88) becomes unity; the equivalent-triode mu, μ_1, then becomes the triode amplification factor μ.

For the pentode, the real part of the transadmittance may be taken as substantially

that of a tetrode, the phase being increased by the increase in transit angle between the screen grid and plate of the pentode as compared with that in the tetrode.

CATHODE-RAY TUBES

Cathode-ray tubes ordinarily employ indirectly heated, oxide-coated cathodes. The electrodes are usually made of stainless steel or nickel. The magnitude of the electron beam is controlled by means of a grid comprising essentially a disk having a single hole. The beam may be focused on the screen by either an electrostatic lens or a magnetic lens. Similarly, the beam or beams may be deflected on the screen by either electric fields or magnetic fields.

In the following sections some of the more important principles of construction and operation will be outlined as they apply to cathode-ray tubes in general, including television tubes for color display. No attempt is made to distinguish between various kinds of cathode-ray tubes; color tubes are not discussed in detail because they are in an evolutionary stage and are likely to remain so for many years.

41. Focusing. In most cases it is easier and more economical to use electric focusing and deflection, as is done in oscillographic and low-price television equipment. Magnetic focusing produces a better spot, and magnetic deflection results in less deflection defocusing. This permits wider angle deflection, hence a shorter tube. Thus, although somewhat more expensive, all-magnetic tubes are at present preferred in high-quality television receivers.

The focusing and deflection of electrons in a cathode-ray tube depend upon the forces exerted on them by the electric and magnetic fields through which they move between cathode and screen. These forces are vectors, and although the quantitative relations that will be given are in scalar notation, it should be understood that the reference axes must be chosen to make this convenience valid. Thus the electric force $f_{(e)}$ is

$$f_{(e)} = -eF$$

where e is the charge and F the field. Taking the perpendicular axes x, y, and z, the forces will be

$$f_x = edV/dx$$
$$f_y = edV/dy$$
$$f_z = edV/dz$$

where f_x is the force in the x direction, $-dV/dx$ is the field in the x direction, etc.

Similarly, the magnetic force is

$$f_{(m)} = Bev$$

where B is the magnetic field perpendicular to the path of the electron and v is the electron velocity. The forces in the three directions will be

$$f_x = e[B_z(\partial y/\partial t) - B_y(\partial z/\partial t)]$$
$$f_y = e[B_x(\partial z/\partial t) - B_z(\partial x/\partial t)]$$
$$f_z = e[B_y(\partial x/\partial t) - B_x(\partial y/\partial t)]$$

where the Bs are the magnetic field components in the subscript directions and the derivatives are the velocity components in the indicated directions.

While the application of these principles to the design of a satisfactory focusing system is a highly specialized problem, the deflection for small angles may readily be derived.

42. Electrostatic deflection, neglecting edge effects, requires that the deflection field be the quotient of the voltage between the plates by the distance between them. The force will be applied during the time $t = l/v$, so that the electrons emerge at an angle given by

$$\tan \phi = (l/2)(F/V_0) \tag{89}$$

where ϕ = angle of deflection
l = effective length of deflection system
F = average field perpendicular to axis
V_0 = beam potential

Neglecting second-order effects, the displacement d of the spot on the screen may be written

$$d \simeq (Ll/2x)(e_d/e_0) \tag{90}$$

where L = distance from center of deflection system to screen
x = separation between deflection plates
e_d = voltage between deflection plates
e_0 = voltage between deflection system and cathode

43. Magnetic Deflection Systems. Here it is assumed that the electrons enter the deflection system along the z axis. With the magnetic field perpendicular to the z axis or axis of the tube, the electrons will have a force applied in the $+y$ direction at first, then in the $+y$ and $-z$ directions as the velocity component in the y direction becomes finite. Therefore, the electrons describe a circular path in a plane perpendicular to the applied magnetic field, so that they emerge at an angle ϕ from the axis given by

$$\phi = (lB_x/V_0)\sqrt{e/2m} \tag{91}$$

where B_x is the field perpendicular to the axis, m is the mass, and e the charge. For practical purposes, the displacement of the spot on the screen will be

$$d \simeq LlB/3.37\sqrt{V_0} \tag{92}$$

where B is the field in gausses.

44. Defocusing Effects. Both systems have inherent defects due to the fringing fields at the edges of the focusing system operating on a beam of finite diameter, the net effect being known as *deflection defocusing.* Some compensation is achieved through careful attention to the shapes of the deflection fields. In general, deflection distortion is minimum when the length of the deflection system is maximum, the limit being to a considerable degree a compromise between distortion, mechanical size, and circuit loading.

45. The Ion Trap Gun. The beam of a cathode-ray tube consists almost entirely of electrons, but there is present in addition ions of various elements in the tube, the most prevalent generally being that of oxygen, O_2^-. Where the ions strike the screen, they cause a reduction in the luminous efficiency of the phosphor. In a system employing electrostatic focusing and deflecting, the ions are focused and deflected in the same manner as are the electrons, since the deflection is independent of mass, as seen by Eq. (90). With the spot continuously moving, the effect of the ions is spread more or less uniformly over the screen so that the decrease in luminous efficiency is not great. However, where the beam is electrostatically focused and magnetically deflected, the ions and electrons reach the deflection system moving together on or near the axis of the tube. The angle of deflection, as shown by Eq. (90), is inversely proportional to the square root of the mass, so that, for instance, an oxygen ion O_2^- is deflected only by an amount $d/\sqrt{32 \times 1{,}850}$, which would be about 0.01 in. when the spot due to the electron beam is displaced 3 in. The result is that in time a dead spot, or *ion burn*, occurs in the center of the screen.

Ion burns are prevented in tubes employing magnetic deflection when an ion trap gun is used and are inhibited by the use of an aluminum screen backing.

In an ion trap gun the beam is first formed, then bent by means of a constant magnetic field. The electrons are then allowed to pass on toward the screen, while the ions, which were substantially undeflected, are captured in the electrodes or on the neck of the bulb. This requires that the tube be provided with a bend in the gun or neck of the bulb.

It is evident, however, that if a magnetic field and an electric field are applied at

right angles to each other at substantially the same location on the axis, the ions may be separated from the beam when the fields are adjusted to produce equal but opposite deflections *for electrons* by Eqs. (90) and (91). Although such an arrangement is not readily achieved in practice, guns have been made in which the electron path does not move far off the axis of the tube and the ions are substantially all removed from the beam.

Table 5

RETMA No.	Sylvania type No.	Material	Fluorescent color	Phosphorescent color	Persistence	Approx avg particle size, μ	Remarks	Application
P-1	160	Zn_2SiO_4:Mn	Green	Green	Medium	2	Small particle size	Oscilloscopes
P-1	161	Zn_2SiO_4:Mn	Green	Green	Medium	5	....	Oscilloscopes
P-1	221	Zn_2SiO_4:Mn	Green	Green	Medium	10	Large particle size	Oscilloscopes
P-2	145	ZnS:Cu	Blue-green	Green	Long	30	....	Oscilloscopes and radar
P-4	CR-40 and 401	ZnS:Ag and (Zn,Cd)S:Ag	White	Blue-white	Medium-short	10	....	Television receivers
P-4	CR-421 and 422	ZnS:Ag and (Zn,Cd)S:Ag	White	Blue-white	Medium-short	10	Silica coated	Television receivers
P-5	135	$CaWO_4$	Blue	Blue	Very short	8	....	Special oscilloscopes (fast photography)
P-7	130	ZnS:Ag	Blue	....	Medium-short	15	Two-layer (cascade)	Radar
	140	(Zn,Cd)S:Cu	Yellow	Yellow	Long	30	Screen	
P-11	132	ZnS:Ag	Blue	Blue	Short	5	....	Oscilloscopes for visual or photographic observation
P-12	144	$(Zn,Mg)F_2$:Mn	Orange	Orange	Medium-long	6	....	Radar
P-13	211	$MgSiO_3$:Mn	Red	Red	Medium	2		
P-14	130	ZnS:Ag	Blue	....	Medium-short	15	Two-layer (cascade)	Radar
	146	(Zn,Cd)S:Cu	Orange	Orange	Long	30	Screen	
P-15	137	ZnO:Zn	Blue-green	Blue-green	Very short	4	....	Flying-spot scanners
P-22	151	$Zn_3(PO_4)_2$:Mn	Red	Red	Medium	3–6	....	Color TV receivers
	161	Zn_2SiO_4:Mn	Green	Green	Medium	3–6		
	131	ZnS:Ag	Blue	Blue	Medium-short	4–8		
	110	(Zn,Cd)S:Ag	Red	Red	Medium-short	10–20		
	170	$CaMg(SiO_3)_2$:Ti	Blue	Blue	Medium-short	3–5		

46. Aluminum-backed Screens. The aluminum screen backing consists essentially of a very thin film of aluminum laid flat over the screen. Its principal purpose is to reflect the light emitted by the phosphor so that it is passed back through the face of the tube where it may be seen, thus improving the over-all efficiency. While the loss in energy of an electron in passing through the aluminum backing may be of the order of a few thousand electron volts, the over-all efficiency of the aluminum-backed screen is greater than the unbacked screen at beam potentials greater than about

5 kv. The aluminum backing also has an inhibiting effect on ion burning. It is produced by volatilizing aluminum metal onto a thin membrane stretched across the phosphor, then removing the membrane by thermal decomposition. The membrane is produced by first placing enough water in the bulb to cover the faceplate, then adding a small amount of nitrocellulose diluted to 3 per cent in amyl acetate; when the solvent evaporates, the water is poured out from under the solidified film.

47. Phosphors. The coating of the screen in a cathode-ray tube is called a ***phosphor*** and the process by which part of the energy of the bombarding electrons is converted to light is called *fluorescence.* When the beam is cut off, the light output due to *phosphorescence* usually decays rapidly at first, then more slowly.

Most of the phosphors used in commercial tubes are sulfides or silicates activated with silver or manganese, although titanium and copper are sometimes used as activators. Table 5 shows the composition and characteristics of phosphors standardized by the Radio-Electronics-Television Manufacturers Association. It should be noted that some phosphors of the same general type, such as the P-22 phosphors used in color TV tubes, are made in several varieties and are sold according to the manufacturer's type numbers; Sylvania type numbers are given in the second column to illustrate commercial practice.

48. Screening. The process of settling the phosphor on the faceplate of the cathode-ray tube is called *screening.* Details of practices vary among manufacturers and among tube types; an outline of the procedure followed for a 21-in. black-and-white TV picture tube follows as an illustration.

1. Place 600 cc of 1 per cent barium acetate in bulb; then add 22 liters of water.
2. Add 1,100 cc of 0.5-normal potassium silicate. Simultaneously add phosphor suspension, consisting of 100 g of phosphor ball-milled in 300 cc of water, then diluted to working consistency.
3. After slight agitation to ensure mixing, the suspension is allowed to settle, the water is poured off, and the bulb is then baked to solidify the screen.

Most phosphors produce maximum light output when the amount of phosphor used is equivalent to about 4 or 5 mg per cm^2, although there are exceptions. The phosphor manufacturer ordinarily is able to provide a recommended figure.

All water used for solutions and suspensions in processing cathode-ray tubes must be either distilled or deionized, as the salts normally occurring in tap water produce undesired effects. Copper salts cause a shift in screen color; iron salts cause degeneration of fluorescence; other compounds contribute to emission failure or the occurrence of gas.

The potassium silicate bonds the phosphor particles together and to the faceplate of the tube. The barium acetate promotes jelling of the silicate.

In tubes operated with beam potentials in excess of 12 kv, it is common practice to use coated phosphors in order to prevent burns due to ion bombardment. The coating process consists in precipitating a thin film of silica on the phosphor particles. This operation is normally performed by the phosphor manufacturer. Silica coating usually results in a decrease in light output of about 5 per cent.

CHAPTER 8

MICROWAVE TUBES

By G. C. Dalman[1]

1. Introduction. The field of microwave tubes today covers a vast and rapidly expanding area, and to expound it fully would be a formidable task. The scope of this chapter has therefore been limited to a concise presentation of the physical picture of microwave-tube behavior and a description of the important present-day microwave-tube types.

In the following paragraphs a discussion of microwave-tube theory encompassing three categories is given: the electron beam of the tube, the tube microwave circuit, and the interaction phenomena which occur between the beam and the circuit. The discussion is limited to those portions of the theory already proved successful, and only enough of the underlying physics will be given to make the physical picture of the operating principles understandable.

2. Energy-conversion Processes. The basic problem in microwave tubes, as with other amplifier, oscillator, and frequency-multiplier tubes, is to convert energy from a d-c power supply into useful r-f energy. In general, this is accomplished by arranging, within the tube, a suitable interaction between an electron beam and an electromagnetic field. In one method commonly used, electrons which have received high kinetic energies by acceleration through a d-c field are caused, by proper design of tube structure and tube circuit, to be decelerated by an electromagnetic field which has been set up by the r-f voltage to be amplified or generated. The energy lost by the electron due to this deceleration is, by conservation of energy, gained by the r-f field. For the ideal case of 100 per cent energy transfer, the electrons are slowed down to zero velocity and all the kinetic energy of the electrons is converted to useful r-f energy.

An alternative method for converting the d-c power-supply energy into useful energy is to arrange an interaction structure in which the electrons of the beam pass through a d-c field without a change in the kinetic energy of the electrons. In this case the d-c potential energy is said to be transferred into useful r-f energy.

All microwave tubes are based on either or both of these types of interaction, and the chief differences among them originate in the particular manner in which the electrons are allowed to interact with the electromagnetic field, the type of electron beam used, and the arrangement of the electromagnetic circuit. In the following paragraphs the important differences are described.

ELECTRON BEAMS

In general, an electron gun is a device for providing a beam of high-velocity electrons arranged in a particular geometrical shape, *e.g.*, cylindrical, ribbon, hollow tube, etc. It is comprised usually of a properly shaped cathode, an anode, and a means for focusing the beam to the desired shape. Usually no r-f field exists in the cathode-anode space. If we classify, however, the cathode-anode region of the magnetron and microwave diode and triode as electron-gun structures, these types are notable exceptions. A large variety of gun structures are used in present-day microwave tubes, the most important of which will be described to illustrate, in a general way, some of the many practical considerations in selecting a beam for a microwave tube. Considered briefly are focusing methods in the accelerating and drift regions of the gun, limitations

[1] Cornell University.

imposed by the electron emitters used, and a summary of some important phenomena observed in electron beams.

3. Cathode Considerations. Most microwave tubes are designed to deliver moderate power (20 mw minimum) or high power (up to 40 megawatts pulsed), and for these tubes the electron beams used are characterized by high-voltage high-current operation. For c-w-operated tubes, beam voltages and beam current densities range from 10^2 to 10^4 volts and 0.050 to 50 amp per cm^2, respectively, and for low-duty-cycle pulse operation, beam voltages and beam current densities as high as 400 kv and 10 amp per cm^2 are used. Important problems are the limitations imposed by the electron-emitting source such as the maximum allowable current density and effects of ion bombardment.

Three types of electron sources for beams are in common use: oxide-coated nickel, pure metal, and thoria. A wide variety of dispenser types are in experimental use. The important characteristics of these electron sources are discussed below.

4. Oxide-coated Cathodes.* Long-life operation can be obtained from c-w-operated oxide cathodes operating up to about 1,500 beam volts; and for pulse operation, provided the pulse duration is kept shorter than the ionization time of the residual gas in the tube, safe operation up to 400 kv is possible. This type of cathode has the widest application, and the methods for producing them are best understood. Its chief limitation is that the safe operating current density is limited to about 0.2 amp per cm^2 maximum for c-w operation and approximately 5 amp per cm^2 maximum for pulse operation at normal operating temperatures (800 to 825°C). A second limitation is the deleterious effect of ion bombardment which occurs for high-duty-cycle operation above 1,500 beam volts. The application of the oxide cathode in microwave-tube beams, however, has been remarkably successful.

5. Pure-metal Cathodes. To permit c-w operation at high voltage (e.g., 5,000 to 10,000 volts) without ion-bombardment damage, an emitting surface consisting of a disk of high-melting-point metal such as tantalum, molybdenum, or tungsten is sometimes used. Because of the high work function of these metals it is impractical to heat the disk to emitting temperatures with conventional heaters of the type used with the oxide-coated types. It is necessary rather to use an auxiliary emitter such as a pure or thoriated-tungsten filament to bombard the opposite face with electrons. The total dissipation of this subsidiary diode is adjusted so that the temperature of the cathode-emitter material reaches the desired value.

While this type of cathode is especially insensitive to ion bombardment, the added complexity of the subsidiary diode structure and power-supply equipment makes it rather unsatisfactory. Thoriated-tungsten cathodes have been proposed but are not in commercial use.

6. Thoria-type Cathode.[1,†] At present there are two types of thoria emitters. The first is an indirectly heated type in which, following the practice of the conventional oxide cathode, a thin coating of powdered thorium oxide is formed on a base metal. Because the normal operating temperature of the cathode is approximately 1600°C, the refractory metals (tungsten, tantalum, or molybdenum) are used as base metals. The cathode operating temperature is much lower than for the pure-metal cathodes, and it is possible to design tungsten heaters to heat the thoria to 1600°C. The heater, however, must operate in the approximate range of 2400 to 2700°C which can cause short heater life due to the high evaporation rate in vacuum of tungsten at these temperatures.

The second type of thoria cathode is directly heated. This cathode is formed by molding a mixture of powdered thorium oxide with powdered tungsten or molybdenum into a suitable form (*e.g.*, cylinder) and firing at about 2000°C. Electrical and mechanical contact is made at each end of the cathode and heating is accomplished by passing a high current directly through the cathode structure.

An important advantage of the thoria cathode is its ability to withstand high-

* See also discussion of G. D. O'Neill in Chap. 7.

† Superior numbers indicate the literature references at the end of this chapter.

voltage ion bombardment. Higher pulse- and c-w-operation-current densities than are possible with conventional oxide cathodes are claimed; sufficient operating experience in commerical tubes, however, has not been accumulated to establish definite values. The thoria-type cathode, although not in widespread use, has some important applications in magnetrons and power tetrodes.

7. Miscellaneous Dispenser-type Cathodes. In recent years another group of experimental emitters[2-6] has been developed with properties intermediate between the standard-type oxide-coated emitter and the pure-metallic emitter. These can be classified roughly as dispenser types. Essentially, they contain a large amount of low-work-function semiconductor material, such as the standard-type oxide coating either uniformly or nonuniformly distributed in a porous-type plug. One surface of the plug faces the vacuum and acts as the electron-emitting source. The porous material used for the plug is generally sintered nickel or tungsten powder. During operation a large amount of active metal (*e.g.*, free barium) is produced and diffuses to the emitting surface, continuously replenishing the active metal which has been evaporated or sputtered. The chief advantages of this cathode, compared with the standard oxide cathode, are that it is much less sensitive to ion bombardment and that c-w-operation current densities of 1 to 5 amp per cm^2 are practical. This type of emitter, however, must operate about 100 to 150°C hotter than the standard oxide cathode. Usually the rate of evaporation of the active metal is high and sometimes undesirable phenomena such as grid emission result.

Improvements in microwave-tube performance can be expected as improvements in the dispenser-type cathodes are made.

Other types of cathode, such as the lanthanum boride cathode, have been proposed but the practical importance of these has not yet been established.

ELECTRON GUNS

8. Principal Gun Types. In broad terms, electron guns can be classified as ***convergent***, *parallel-flow*, and *divergent* types. Convergent guns find their application where very high values of beam current densities are required as in many klystrons and traveling-wave tubes. They are used where the beam current densities are greater than the allowable safe values available directly from the cathode surface. The parallel-flow type, simpler in structure, is used where the current densities available directly from the cathode are adequate, as in microwave diodes and triodes and some klystrons and traveling-wave tubes. Although not in as widespread use as the convergent type, improvements in high-current-density cathodes, such as the dispenser type, will increase the importance of parallel-flow guns. The divergent gun at the present time has no commercial application.

For conciseness, the discussion has been limited here to the convergent-type gun, the most important of the three types. The discussion should serve to illustrate in a general way the important beam-operating characteristics and problems encountered. Taken up are the basic structures of the principal gun types, a discussion of the method of focusing the beam in the cathode-anode (accelerating) region, and methods for maintaining the beam shape in the field-free "drift" space. A limited discussion of some special beam phenomena is also included.

9. Gun Structures. The most common type of convergent gun, usually called a "Pierce gun,"[7] is illustrated schematically in Fig. 1. It consists of a cathode and anode which are portions of a spherical diode, and a focusing structure. For large values of the semiangle θ, the anode often takes the form of a spherically shaped tungsten mesh, and for low semiangles, a flat-anode grid is often satisfactory. Gun microperveances* in commercial-type tubes are as high as 15, and gun semiangles as high as $\theta = 35°$ have been used. An important feature of the Pierce gun is that the current density is uniform across the cathode and anode surfaces.

* Microperveance is 10^6 times the beam current (amperes) divided by the $3/2$ power of the beam voltage (volts$^{3/2}$).

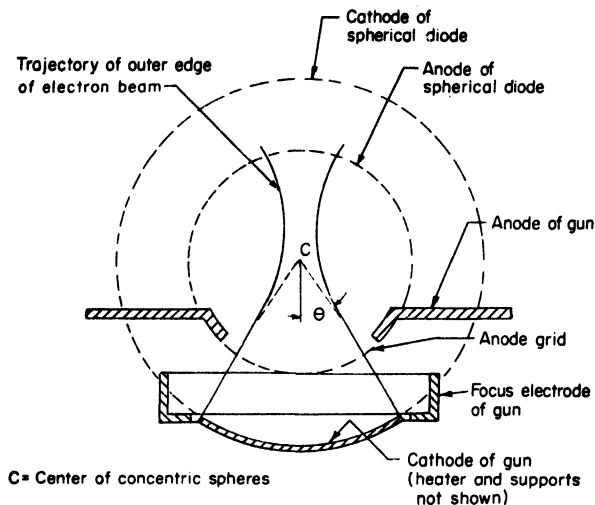



FIG. 1. Schematic diagram of a "Pierce" gun.

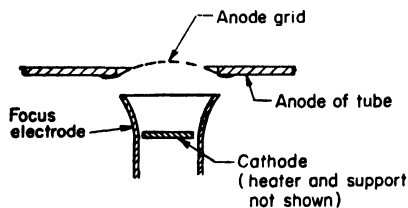



FIG. 2. Empirically designed gun (2K25). FIG. 3. Gridless-anode gun.

Another convergent-type gun which is of empirical design is shown in Fig. 2. This gun was widely used in some of the early reflex klystrons.[8]

In many applications it is desirable to eliminate the anode grid. Several gridless-anode guns have proved successful, and these are shown in Figs. 3, 4, and 5. The gun of Fig. 3 is a gridless-anode type designed[9,10] for uniform cathode-current density. Figures 4 and 5 show two high-angle empirically designed guns.[11] The gun of Fig. 3 is widely used in klystron and traveling-wave tubes, and the principal use of those illustrated in Figs. 4 and 5 has been in reflex klystrons.

FIG. 4. High-angle gridless-anode gun.

10. Focusing Methods in the Cathode-Anode Space. The focusing structures for microwave tubes of early design were determined empirically. While some of these guns have excellent characteristics, the design is determined by trial and error and is extremely laborious. These guns often suffer from the defect of nonuniform cathode-current densities. However, the guns of Figs. 2, 4, and 5 have proved very satisfactory. Once a satisfactory design is established, these guns can be scaled for higher- or lower-voltage operation.[12]

One of the most important advances in gun-design techniques has been suggested by Pierce,[7] who proposed that the focusing structure be designed to provide boundary

conditions at the outer edge of the beam which is identical to that known to exist[13] in the spherical diode shown dotted in Fig. 1. The electrons of the beam will then flow in the gun in the radial direction only as in the spherical diode. The design of the focusing structure is determined by placing a scale model of the gun in an electrolytic tank, and the shape of the focusing structure is adjusted experimentally to provide the proper boundary conditions at the beam edge.[7,13] The most easily fabricated shape of a wide variety usually possible is selected for actual use.

Transmission efficiency 95%
Perveance .... 3 microamp/volt $^{3/2}$
Current density multiplication 75
Heater power .... 3.9 watts

Beam voltage (volts)	400	500	600
Cathode current (ma)	24	33	45
Current density at cathode (ma/cm^2)	180	250	340
Current density at beam minimum (amps/cm^2)	14	19	25

FIG. 5. Heil gun. (*Bell System Tech. J.*)

11. Beam-focusing Methods. The next problem is to maintain the desired beam shape along the axis of a traveling-wave tube, through klystron cavity grids, etc., so that proper interaction between the beam and the circuit can occur. Four methods of controlling the beam shape are in use: *magnetic focusing* in which a magnetic field overcomes spreading effects due to the space-charge forces within the beam; *ion focusing* in which positive ions neutralize the space charge of the beam; *electrostatic focusing* in which space-charge spreading is corrected by a series of converging lenses; and *space-charge focusing* in which the beam follows the space-charge spread curve.

FIG. 6. Space-charge-focused beam in three-cavity amplifier.

The term space-charge focusing is used loosely since the beam is actually allowed to travel "unfocused" and the circuit is fitted to the beam. Figure 6 shows an example[14] of a tube employing a space-charge-focused beam. In this case the beam was designed so that the minimum beam diameter occurs approximately at the center of the tube. Beyond the location of the minimum beam diameter, the space charge in the beam causes the beam to diverge, making it necessary to increase the diameter of the third cavity grid. This tube is pulse-operated and the pulse duration is less than the ionization time of the residual gas within the tube.

Figure 7 shows an example of the tube employing ion focusing.[14] Ions formed by electron bombardment of the residual gas molecules in the tube are trapped within the beam neutralizing the space charge. Thus the beam travels beyond the plane of the

anode unaffected by space-charge spreading. While ion focusing has been successful in many microwave tubes, in some applications the ions introduce some undesirable second-order phenomena such as ion noise and plasma oscillations.

The third type of beam focusing utilizes a magnetic field. Early magnetically focused beams were focused by immersing the entire tube, including the gun structure, in a uniform axial magnetic field. The focusing action is based on the fact that electrons tend to follow magnetic-field lines. The method is limited to parallel-flow guns and has the disadvantage that a very large magnetic field, theoretically infinite, is required to prevent the beam from spreading. An important refinement in magnetic focusing[7,15] involves providing a particular magnetic field so that the space-charge repulsion forces and the centrifugal force of rotation are just balanced against the centripetal magnetic force on the electrons. If the beam enters the magnetic-field region properly, the outer-edge electrons spiral at constant diameter down the tube and the beam is said to have *Brillouin flow*.

FIG. 7. Ion-focused beam in a three-cavity amplifier with gridded gaps (SAS-28).

The considerations involved in the proper design of such a beam using a uniform axial field are illustrated in Fig. 8. In the ideal case, the beam is injected abruptly in a constant magnetic field from an electrostatic-type electron gun which is excluded from the magnetic field. In the actual case, however, the beam is injected gradually into the magnetic field, some magnetic flux links the cathode, the magnetic field is not completely uniform along the tube axis, and the beam may not enter the magnetic field with zero radial velocity. Wang[15] has made a complete study of this problem and with the aid of radial and axial force equations has shown how Brillouin flow can be

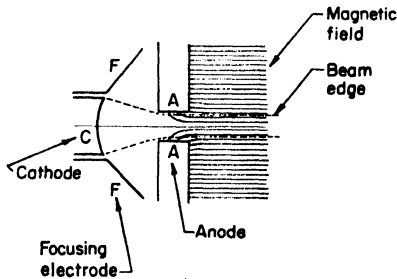



FIG. 8. Magnetically focused beam (shaded over).

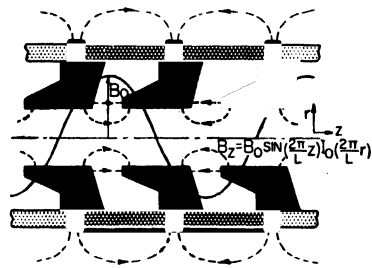



FIG. 9. Periodic focusing structure.

obtained successfully. If the proper magnetic field and entrance conditions are not provided, the envelope of the beam will be scalloped in shape and beam interception or poor interaction usually results.

An important disadvantage to the use of magnetic focusing is the heavy and bulky permanent magnets or electromagnets required. These defects can be avoided by the use of periodic magnetic focusing.[17] Small permanent magnets, as in Fig. 9, are stacked to produce alternating-polarity magnetic fields, periodic in space. The principal application of periodic focusing has been in traveling-wave tubes.

The feasibility of periodic electrostatic focusing is definite, although no applications

can be found in commercial tubes to date. The demand for light-weight tubes of the traveling-wave types will probably lead to the development of a series of commercially available tubes.

In the following paragraphs are brief discussions of some important special beam structures and focusing methods.

12. Hollow Beams. Often the beam dimension and the physical structure of the r-f circuit of the tube are such that the interaction occurs mostly at the outer periphery of the beam, *e.g.*, as in some gridless cavity-gap klystrons, traveling-wave tubes, and

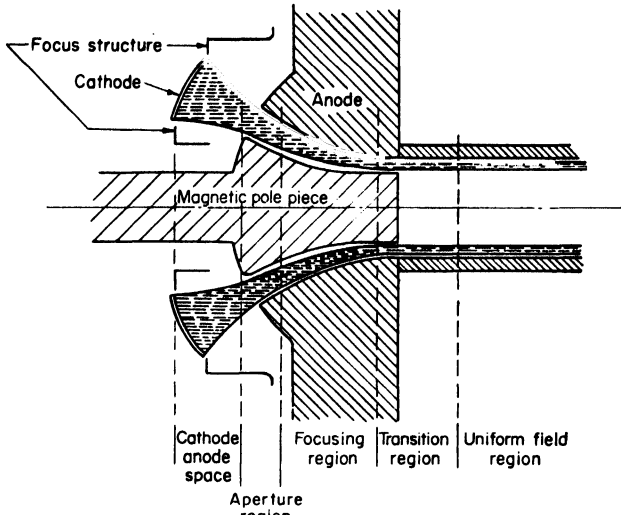



FIG. 10. Sperry hollow-beam cathode.

backward-wave oscillators. Because the center of the beam does not contribute to the output power, the use of a hollow beam can make significant improvements in tube efficiency. The simplest method for obtaining a hollow beam is to omit the oxide coating from the center of the cathode of a parallel-flow cylindrical-type immersed gun, and this type has found application in backward-wave oscillator tubes. Some progress on convergent-type hollow beams has been made by Wang and Bol,[16] but considerable work remains. An experimental hollow-beam gun having a microperveance of 10 is shown in Fig. 10.

13. Beams for Crossed-field Tubes. The gun structure and electron beam for crossed-field tubes are shown schematically in Fig. 11. A beam of electrons, properly focused in the gun region, travels perpendicular to two mutually perpendicular d-c electric and magnetic fields. A simple and systematic design procedure remains to be developed for beams for the crossed-field-tube types.

FIG. 11. Beam for crossed-field tube.

14. Modulated Beams. In many microwave-tube applications it is desired to pulse the tube to obtain a modulated signal, often of very high powers. A method often used for tubes employing non-ion-focused guns is to pulse the beam voltage directly. This has the disadvantage of requiring a high-power modulator. To overcome this difficulty an auxiliary anode grid is placed directly in the beam close to the cathode. Because of its close proximity to the cathode the beam current can be controlled by a relatively low-power auxiliary anode grid modulator. If an auxil-

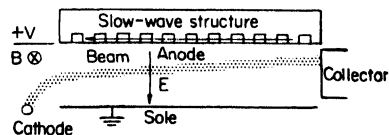
iary anode of the type shown in Fig. 12 is used, little or no power is required of the modulator.

15. Low-noise Beams. An important consideration in the development of low-noise tubes is the noise properties of the beam. In effect, the noise sources in the cathode produce a noise current in the beam. This current is amplified along with the signal current introduced on the beam by the interaction circuit so that the output consists of both amplified signal and noise. An important indirect cause of this noise is the velocity spread (Maxwellian distribution) of the electrons emitted from the cathode. As a result of the interaction effects in the gun and beam region[7,17,18,19] the velocity fluctuation at the cathode becomes a current fluctuation at the anode planes because of electron-bunching action (see Interaction Theory to follow). Two other sources of noise exist: current fluctuations at the cathode surface and interception noise. The former is believed of lesser importance than the velocity fluctuation but plays an important role in determining the minimum noise figure of a microwave amplifier. *Interception noise* results when a beam strikes a part of the tube wall, *e.g.*, the gun anode.

FIG. 12. Gun with modulation anode. (*Eitel-McCullough.*)

In tubes where beam interception can be minimized, the dimensions and voltages can be arranged[18,19,20] so that the noise current at the beginning of the interaction circuit is a minimum and, consequently, the noise figure is a minimum. These guns have been developed to the point where low-noise traveling-wave tubes are now commercially available having an average noise figure of 6.5 db, over a 2,700- to 3,500-Mc band (RCA type 6861). In tube types where interception is high, such as gridded-gap klystrons, interception noise is excessive and very low noise figures are impossible to attain. For large beam-interception ratios, the beam noise is essentially equal to full shot noise.[12] Noise figures less than 4 db can be expected in the near future.

CIRCUITS FOR MICROWAVE TUBES

The circuits for microwave tubes fall into two important categories: interaction circuits and coupling circuits. Coupling circuits are devices which provide an r-f match between the input or output terminal of the tube and the tube-interaction circuit. The tube-interaction circuit is a structure which is arranged so that an electromagnetic wave on the circuit will have a component of electric field in the space occupied by the beam.

16. Interaction Structures; Resonant Cavities. The simplest form of microwave-interaction circuit is the reentrant-type cavity.[21,22] The cavity structure often takes the form of a figure of revolution about a center line such as shown schematically in Fig. 13*a*. Cavities of this type have high Qs (approximately 50 to 10,000) so that relatively small amounts of r-f power fed into the cavity will cause a high r-f voltage to appear across the cavity gap. If the cavity-gap width is made small enough so that the transit time for an electron passing through the gap is not too great, then the cavity is a useful type of interaction circuit.

A variety of similar cavity-interaction structures are also used such as doubly reentrant types and reentrant wave guides. The particular type chosen in a practical case is usually determined by factors such as band width, power-handling capabilities, and ease of manufacture.

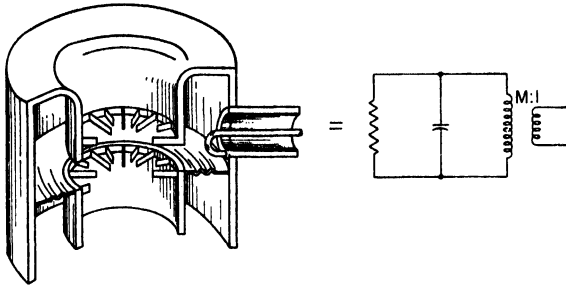
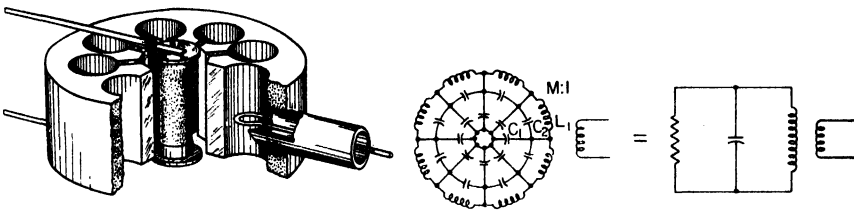
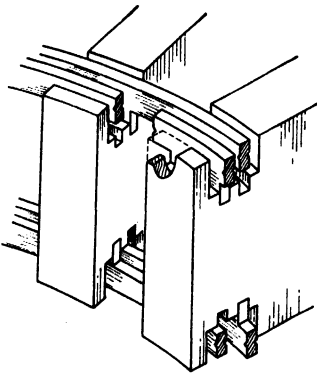
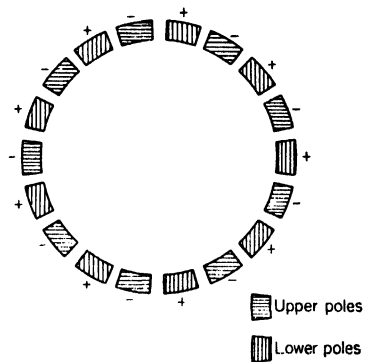
Another important cavity-interaction circuit is the magnetron circuit shown schematically in Fig. 13*b*. This circuit consists of a group of coupled resonant cavities arranged in a ring. The behavior of the circuit can be described approximately by

(a) Reentrant cavity

(b) Magnetron cavity

(c) Strapped magnetron cavity

(d) Interdigital magnetron

FIG. 13. Resonant-cavity interaction structures.

the equivalent circuit shown in the figure. When excited at the resonant frequency two waves travel in opposite directions with a phase velocity which is some fraction of the speed of light. If an electron beam is arranged to travel in a circular motion at a speed approximately equal to the phase velocity, interaction will occur with the wave traveling in synchronism with the beam.

There are many possible operating modes of the magnetron cavity: the most commonly used is the mode in which the r-f fields across the gaps of adjacent cavities are π radians out of phase. This mode is called the π mode. Several methods are

employed to suppress other possible cavity modes, the most common of which is to connect alternate cavity gaps to two separate rings, as illustrated in Fig. 13c. Similar types of magnetron cavity structures are also used in commercial-type magnetrons.[23,24] A linear array of cavities is sometimes used in magnetron amplifiers.

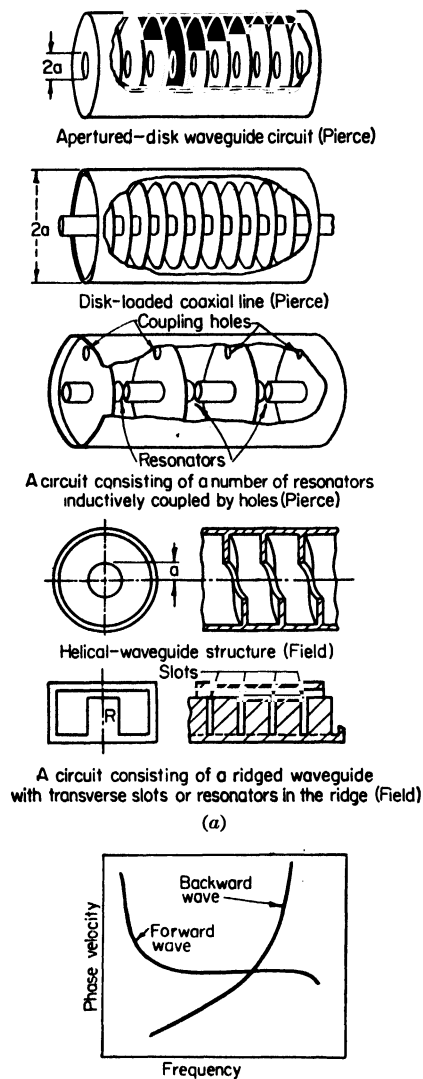
FIG. 14. (a) Slow-wave structures. (b) Phase-velocity variation in the forward- and backward-wave circuits.

The cavity-interaction structures are resonant devices and standard methods are used for tuning: mechanical adjustments of the cavity gap, introduction of a plunger into the cavity to change the volume, susceptance tuning by means of a coupled cavity (*i.e.*, by frequency pulling), etc.

Interdigital Magnetron Resonant Structure. Illustrated schematically in Fig. 13d is one arrangement of an interdigital magnetron resonant structure comprised of two sets of intermeshed fingers, each set being connected to opposite poles of a single-cavity resonator not shown. Tuning is accomplished by adjustment of the cavity-resonator frequency. The fingers of the structure are used as the magnetron anode, and the cathode, also not shown in the figure, is placed at the axis of the anode.

17. Nonresonant Circuits. These consist of structures capable of propagating electromagnetic energy in several possible modes, one of which has a phase velocity slow enough to travel in synchronism with an electron beam passing through or adjacent to the structure. Useful structures are those having a component of electric field in the direction of the beam path. Thus over a long distance, the electron motion will be influenced by the field and vice versa. Many structures are capable of supporting slow waves, a few of which are illustrated in Fig. 14. The most widely used of these circuits is the helix (Fig. 17). It is clear from an inspection of the helix structure that, although the electromagnetic energy travels around the helix with the speed of light, the velocity of propagation of the axial fringe field is considerably less.

When these circuits are used in amplifiers, an attenuator is usually placed on the circuit near the input end to prevent oscillation due to reflected waves near the output. Usually the amount of attenuation is greater than the gain of the amplifier.

The helix-type structure is especially suitable for wide-band tubes since it is nondispersive, *i.e.*, the phase velocity does not change rapidly with frequency as illustrated in Fig. 14b. Other structures have been used in which the phase velocity is slowed

down by periodic discontinuities such as apertured disks or bends in the wave guides. These structures, being more dispersive, have a narrow frequency band and find application, because of their greater heat-dissipating ability, in very-high-power traveling-wave tubes.

18. Beam Circuit Figure of Merit. An important parameter of a microwave circuit is the ratio of the effective circuit resistance to the d-c beam resistance. For the case where the beam continuously interacts with the circuit fields (*e.g.*, for a helix) a merit factor, often called the *Pierce C factor*,[25] is defined as

$$C = \left(\frac{K}{4R_0}\right)^{1/3}$$

where R_0 is the ratio of the beam voltage and beam current and K is the effective impedance of the helix defined by Pierce as

$$K = \frac{E^2}{2\beta^2 P}$$

where E is the r-f field acting on the electron beam, β is the phase constant which is the ratio of the angular frequency ω and the average beam velocity v_0, and P is the r-f power propagating down the helix.

A figure of merit for resonant-cavity devices is

$$\gamma = \mu^2 \frac{R_0}{R}$$

where μ is the gap-coupling coefficient, R is the equivalent resistance appearing across the cavity gap due to internal cavity losses, beam loading, etc., and R_0 is the d-c beam resistance. If d is the cavity-gap spacing, the gap-coupling coefficient for gridded cavities is given by

$$\mu = \frac{\sin\left(\frac{1}{2}\omega d/v_0\right)}{\left(\frac{1}{2}\omega d/v_0\right)}$$

19. Coupling Structures. The problem of coupling the tube circuit to an external transmission line is essentially one of microwave-circuit design. Since traveling-wave tubes are usually designed to operate over a 2:1 frequency range, the design problem of obtaining a match (*e.g.*, VSWR < 1.5) is difficult, although realizable. While resonant-structure devices usually have a mechanical tuning range which is considerably smaller than for the traveling-wave tube, the matching problem is more critical since it is often affected by the presence of the electron beam (*e.g.*, the reflex klystron). In this case the matching structure is designed for proper match with the tube operating.

INTERACTION THEORY

Because of the extensive nature of microwave-tube interaction theory, only a brief outline of the principal theories can be given. They are presented here to give the engineer some idea of the theory underlying microwave-tube behavior. All interaction theories start from Maxwell's equations and the Lorentz force equation. The theories differ in detail according to the assumptions which have been made to simplify the mathematical analysis. For example, the assumption of small-signal operation permits an enormous simplification of the mathematical analysis as compared with the more general large-signal case. Other assumptions such as negligible space charge or parallel-plane tube geometry result in further simplification. While much has been accomplished, the theories developed are essentially guides in predicting the actual operation of the tube, and much work remains in this field.

20. Vacuum-tube Network Theory. A classic example of the small-signal analysis of vacuum tubes is the Llewellyn-Peterson[26] vacuum-tube network theory for a system

of parallel planes. Such a system is illustrated in Fig. 15 as an array of planes *a*, *b*, *c*, etc. Currents J_1, J_2, etc., flow in regions 1, 2, 3, etc. The analysis of such a system starts with

$$\Delta \cdot \bar{E} = \frac{\partial \bar{E}}{\partial x} = \frac{\rho}{\epsilon_0} \tag{1}$$

$$\bar{J} = \rho \bar{v} + \epsilon_0 \frac{\partial \bar{E}}{\partial t} \tag{2}$$

$$\bar{F} = m\bar{a} = -e\bar{E} \tag{3}$$

where $\bar{E}$ = electric field
$\bar{J}$ = current density
$\bar{F}$ = force on electron
$\bar{a}$ = acceleration of electron
ρ = charge density
ϵ_0 = dielectric constant of free space
x = distance measured with respect to some reference
$\bar{v}$ = velocity of the electron
t = time
e and m = charge and mass of an electron

Equations (1) and (2) can be combined to eliminate ρ so that

$$J = \epsilon_0 \left(\frac{\partial E}{\partial x} v + \frac{\partial E}{\partial t} \right) = \epsilon_0 \frac{dE}{dt} \tag{4}$$

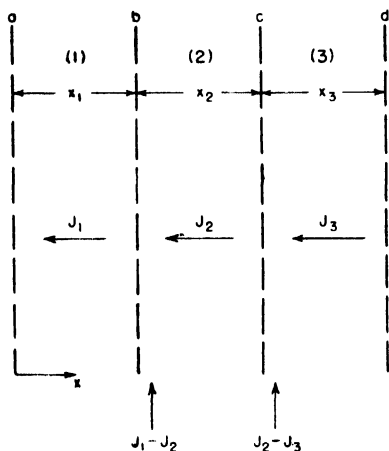
where all vector quantities have x-directed components only. Differentiating Eq. (3) with respect to time and combining with Eq. (4) we have the important relation

$$\frac{da}{dt} = \frac{d^3x}{dt^3} = \frac{e}{m\epsilon_0} J \tag{5}$$

FIG. 15. Parallel-plane vacuum-tube system.

This equation can be solved[27] by assuming that quantities have a d-c component plus a small a-c component. For example, associated with an applied voltage consisting of a d-c component and a small a-c component are quantities such as current density, acceleration, velocity, and transit time which also have a d-c component and a small a-c component. Equation (5) is solved assuming $J = J_0 + J_1 \sin \omega t$ and the voltage required to produce this assumed current is determined after a considerable but straightforward mathematical manipulation. The result of this analysis is summarized in the Llewellyn and Peterson[26] equation:

$$\begin{aligned} V_b - V_a &= AJ + Bq_a + Cv_a \\ q_b &= DJ + Eq_a + Fv_a \\ v_b &= GJ + Hq_a + Iv_a \end{aligned} \tag{6}$$

where the coefficients A through I listed in Table 1 depend on the d-c voltages applied, the frequency, and the spacing between planes, and V = a-c voltage on the plane with respect to ground, J = a-c current density, q = a-c conduction current density, and v = a-c velocity of the electron. Subscripts a and b refer to the particular planes of the array. An inspection of Eq. (6) shows that the conditions at plane b are expressed in terms of the initial conditions at plane *a*.

Equation (6) is a remarkable set of equations in that it represents the complete solu-

tion of the small-signal case of parallel planes. From it, to cite only a few examples, can be derived the equivalent circuit of the microwave diode, triode, tetrode, two- and three-cavity klystrons, etc.[26,27,28]

21. Ballistic Theory. Many microwave tubes are operated at high beam voltages and low current densities. In these cases, the space-charge term of Eq. (1) can be neglected entirely, and one need consider only the conduction-current term in Eq. (2). This is equivalent to assuming the electron to be a simple charged mass whose motion is dependent only on the electric and magnetic fields. There results a considerable simplification of the analysis and in some cases, as for the klystron tube, a quasi large-signal theory can be developed.[12,13,28,29]

Table 1. Vacuum-tube Network Coefficients

$$A = \frac{1}{\epsilon_0}(u_a + u_b)\frac{T^2}{2}\frac{1}{\beta}\left[1 - \frac{\zeta}{3}\left(1 - \frac{12S}{\beta^3}\right)\right]$$

$$B = \frac{1}{\epsilon_0}\frac{T^2}{\beta^3}[u_a(P - \beta Q) - u_b P + \zeta(u_a + u_b)P]$$

$$C = -\frac{1}{\eta}2\zeta(u_a + u_b)\frac{P}{\beta^2}$$

$$D = 2\zeta\left(\frac{u_a + u_b}{u_b}\right)\frac{P}{\beta^2}$$

$$E = \frac{1}{u_b}[u_b - \zeta(u_a + u_b)]e^{-\beta}$$

$$F = \frac{\epsilon_0}{\eta}\frac{2\zeta}{T^2}\left(\frac{u_a + u_b}{u_b}\right)\beta e^{-\beta}$$

$$G = \frac{-\eta}{\epsilon_0}\frac{T^2}{\beta^3}\frac{1}{u_b}[u_b(P - \beta Q) - u_a P + \zeta(u_a + u_b)P]$$

$$H = -\frac{\eta}{\epsilon_0}\frac{T^2}{2}\left(\frac{u_a + u_b}{u_b}\right)(1 - \zeta)\frac{e^{-\beta}}{\beta}$$

$$I = \frac{1}{u_b}[u_a - \zeta(u_a + u_b)]e^{-\beta}$$

where $P = 1 - (1 + \beta)e^{-\beta} \doteq \beta^2/2 - \beta^3/3 + \beta^4/8 + \cdots$
$Q = 1 - e^{-\beta} \doteq \beta - \beta^2/2 + \beta^3/6 - \beta^4/24 + \cdots$
$S = 2 - \beta - (2 + \beta)e^{-\beta} \doteq -\beta^3/6 + \beta^4/12 - \beta^5/40 + \beta^6/180 + \cdots$
$\beta = j\theta = j\omega T$

D-c Equations

Limiting current density:

$$J_m = 2.33 \times 10^{-6}\frac{[(V_{D_a})^{1/2} + (V_{D_b})^{1/2}]^3}{d^2}$$

Definition of space-charge factor ζ:

$$\frac{J_0}{J_m} = \frac{9}{4}\zeta\left(1 - \frac{\zeta}{3}\right)^2$$

Transit time:

$$T = \frac{T_0}{1 - \zeta/3}$$

where $T_0 = \dfrac{2d}{u_a + u_b}$ = transit time in absence of space charge

D-c velocity:

$$u = (2\eta V_D)^{1/2}$$

Distance equation:

$$x = \left(1 - \frac{\zeta}{3}\right)(u_a + u_b)\frac{T}{2}$$

where $\eta = e/m = 1.76 \times 10^{11}$ coulombs/kg
J_0 = density of d.c.
$\epsilon = 1/(36\pi \times 10^9)$ farads/m
V_D = d-c potential
$\eta/\epsilon = 2 \times 10^{22}$

As an illustration of this theory assume Fig. 15 to be a schematic representation of a klystron in which planes a, b and c, d are closely spaced cavity grids separated by the relatively long field-free drift region (2). If a small a-c signal is applied to the first gap, an integration of Eq. (3) will show that the velocity of the electrons leaving the gap is modulated and given by

$$v = v_0 \left(1 + \frac{\mu\alpha}{2} \sin \omega t\right) \tag{7}$$

where v_0 is the velocity of the unmodulated beam and α is the ratio of the peak a-c voltage to the d-c beam voltage, μ is the gap-coupling coefficient previously defined, and ω is the angular frequency of the gap voltage. These electrons are injected into the field-free drift region and the arrival time t_2 of a particular electron at the second cavity gap is dependent on its instantaneous velocity when it left the cavity. If L is the length of the drift space and t_1 is the value of the electron departure time, then for small values of $\frac{1}{2}\mu\alpha$,

$$t_2 = t_1 + \frac{L}{v} \approx t_1 + t_0 \left(1 - \frac{\mu\alpha}{2} \sin \omega t_1\right) \tag{8}$$

where $t_0 = L/v_0$. Since $\theta = \omega t$, this can be written in terms of transit angles as

$$\theta_2 = \theta_1 + \theta_0 - x \sin \theta_1 \tag{9}$$

where $x = \frac{1}{2}\mu\alpha\,\theta_0$, the bunching parameter.

The next important step in the ballistic theory is the application of the continuity-of-charge equation. This states that the total charge arriving at the second cavity gap in the time interval dt_2 consists of the sum of all the charges leaving the first cavity gap in the corresponding intervals dt_1. This is equivalent to

$$I_2\, dt_2 = \Sigma I_1\, dt_1 \tag{10}$$

where I_1 is the d-c beam current at the first gap and I_2 is the beam current at the second gap. From Eq. (8) we have

$$\frac{dt_2}{dt_1} = 1 - x \cos \omega t \tag{11}$$

which shows that for a particular dt_2 there are either one or three corresponding values of dt_1. Thus

$$\frac{I_2}{I_1} = \sum \frac{1}{|1 - x \cos \theta_1|} \tag{12}$$

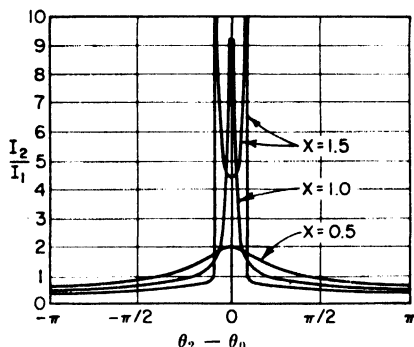



FIG. 16. Plot of I_2/I_1 vs. $(\theta_2 - \theta_0)$.

If various values of θ_2 are assumed, the corresponding values of θ_1 can be calculated from Eq. (9) and substituted in Eq. (12) to obtain I_2. A plot of I_2/I_1 vs. $(\theta_2 - \theta_0)$ for one cycle is shown in Fig. 16. If a Fourier analysis of this curve is taken,[12,13,28,29] the current at the second gap is

$$I_2 = I_1 \left[1 + 2J_n(nx) \sin (n\omega t)\right] \tag{13}$$

where I_1 is the d-c term and n is the harmonic frequency considered. The fraction of this current actually induced in the cavity gap is μ, so that, if the shunt resistance R_{sh} of the second cavity is known, the peak value of the a-c cavity-gap voltage can be calculated as $2I_1J_n(nx)\mu R_{sh}$ and hence the transconductance, voltage gain, etc., can be calculated. These results are identical to those predicted using the vacuum-tube network equations except for the $2J_n(nx)$ term. In a similar fashion, other tube types can be analyzed if the simplifying assumptions made apply.

22. Linear-beam Theory. Clearly the advantage of the vacuum-tube network theory for parallel planes is that it describes completely the small-signal characteristics of a large number of tubes. For tubes having linear beams, *i.e.*, in which the electron beam has a predominantly large velocity component in a particular direction, a similar unified theory has been proposed by Wang.[30] To achieve different interaction functions, various r-f structures are usually built around the beam such as a helix or cavity arrays. Instead of considering a special theory as each tube is developed, a general one is developed which permits a solution of specific tubes with the least amount of modification. It is possible then, with the aid of electronic-computer devices, to determine the characteristics of a wide variety of structures in a relatively short time.

In special cases, closed-form solutions can be obtained, and as an illustration, Pierce's analysis[25] of the traveling-wave tube will be reviewed briefly. In his simplified analysis, a circuit is assumed (*e.g.*, a helix) which has one active mode with a phase velocity near the electron speed. It is further assumed that all a-c quantities vary as exp $(j\omega t + \Gamma z)$, where Γ is the propagation constant $\alpha + j\beta$ of the active

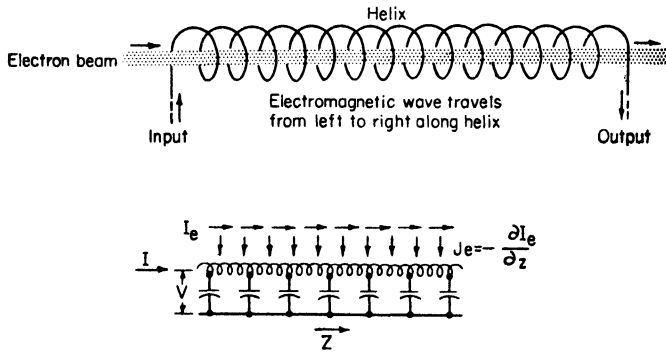



FIG. 17. Equivalent circuit of the traveling-wave tube.

circuit and z is the distance along the length of the tube. In the simple case of lossless circuit $\Gamma = \Gamma_0 = j\beta$ with no beam, where $\beta = \omega/u_0$, the phase constant of the slow wave. If we know β and the effective impedance K of the circuit[25] previously defined, an equivalent distributed active circuit can be drawn having a series reactance X and shunt susceptance B related to β and K according to

$$K = \sqrt{\frac{X}{B}} \tag{14}$$

$$\Gamma_0 = j\beta = j\sqrt{XB} \tag{15}$$

The presence of a beam traveling with approximately the same velocity as the wave is assumed not to change X and B but, because of the bunching in the beam, impresses a current per unit length J on the equivalent circuit. The telegrapher's equations therefore take the form

$$\frac{\partial I}{\partial z} = -jBV + J \tag{16}$$

$$\frac{\partial V}{\partial z} = -jXI \tag{17}$$

where I and V are the equivalent current and voltage on the circuit. If i is the a-c conduction current in the beam, the equivalent circuit in the beam is as shown schematically in Fig. 17. Since all a-c quantities vary as exp $(j\omega t + \Gamma z)$, the solution of

V in terms of i is

$$V = -j \frac{\Gamma\Gamma_0}{\Gamma^2 - \Gamma_0^2} i \tag{18}$$

which gives the voltage (or electric field) on the equivalent circuit in terms of an assumed impressed current i.

Pierce[25] next calculates the conduction current in the beam produced by an assumed voltage V. Since conduction current is the product of the electron velocity and space-charge density, the a-c conduction current is approximately

$$i = \rho_0 v + v_0 \rho \tag{19}$$

where the product of the a-c terms has been neglected for the small-signal case. In Eq. (19) the quantities with the subscripts refer to the d-c components. By a simple application of Newton's force equation and the equation of the conservation of charge v and ρ can be expressed in terms of V. From the force equation

$$\frac{dv}{dt} = \frac{\partial v}{\partial t} + \frac{\partial v}{\partial z} v = \frac{e}{m} \frac{\partial V}{\partial z} \tag{20}$$

or, since exponential solutions are assumed,

$$v = -\frac{e/m\Gamma}{v_0(j\beta - \Gamma)} V \tag{21}$$

From the equation of conservation of charge

$$\nabla \cdot i = \frac{\partial i}{\partial z} = \frac{-\partial \rho}{\partial t} \tag{22}$$

or

$$\rho = -j \frac{\Gamma i}{\omega} \tag{23}$$

Equations (21), (23), and (19) are then arranged as

$$i = j \frac{I_0}{V_0} \frac{\beta\Gamma}{(j\beta - \Gamma)^2} V \tag{24}$$

where I_0 and V_0 are the d-c beam current and voltage.

The final step is to find the values of Γ which satisfy both Eqs. (18) and (24). This is done by equating the two and solving for Γ. To simplify the analysis it is assumed Γ is only slightly affected by the presence of the beam so that

$$-\Gamma = -\Gamma_0 + \zeta \tag{25}$$

where $|\Gamma_0| \gg |\zeta|$. Carrying out the solution the propagation constants are

$$\Gamma_1 = \beta \left[-\frac{\sqrt{3}}{2} C + j\left(1 + \frac{C}{2}\right) \right] \tag{26}$$

$$\Gamma_2 = \beta \left[\frac{3}{2} C + j\left(1 + \frac{C}{2}\right) \right] \tag{27}$$

$$\Gamma_3 = j\beta(1 - C) \tag{28}$$

These are three forward waves. A similar analysis will show a fourth propagation constant of a wave whose direction of propagation is opposite to the direction of the

beam:

$$\Gamma_4 = -j\beta\left(1 - \frac{C^3}{4}\right) \tag{29}$$

Equation (26) shows that one of the three forward waves has a negative attenuation constant. Since the waves vary as exp $(-\Gamma_z)$, the voltage gain of the growing wave of a tube expressed in decibels is $47.3CN$, where C is the merit factor previously defined and N is the length of the tube expressed in wavelengths. The amplitudes of the three forward waves at the input are equal to one-third the impressed signal voltage[25] so that the over-all gain of the tube is

$$\text{Voltage gain (db)} = -9.54 + 47.3CN \tag{30}$$

PRINCIPAL MICROWAVE TUBES

Microwave tubes are used mainly in the 1- to 100-kilomegacycle frequency range. To cover this range a wide variety of tubes have been developed, each with its particular advantage and limitation. The choice of a tube for a particular application will include such considerations as tube efficiency, gain, noise, sensitivity to load, and other electrical characteristics; tube mechanical characteristics such as size, weight, and

Table 2. Family of Microwave Generators and Amplifiers*

I. Microwave triodes and tetrodes
  A. Low-power oscillators and amplifiers
    1. Coaxial electrode structures (pencil and crayon types)
    2. Planar, disk-sealed tubes (lighthouse and rocket types)
  B. High-power oscillators and amplifiers: liquid-cooled, coaxial-structure triodes and tetrodes (RCA "superpower" tubes)
II. Axial beam cavity (lumped circuit) modulation tubes
  A. Single cavity
    1. One transit: monotron†
    2. Two transits: reflex klystron; retarding field oscillator
  B. Multiple cavities: klystron amplifier; electron-stream parametric amplifier
III. Axial beam, traveling-wave (distrib. circuit) modulation tubes
  A. Forward wave
    1. One beam, one circuit: traveling-wave tube (TWT)
    2. Two beams: double stream amplifier†
  B. Backward wave: backward-wave oscillator (O-BWO) or O-type carcinotron; backward-wave amplifier (O-BWA)
  C. Beam-type traveling-wave parametric amplifier
IV. Crossed-field tubes
  A. Resonant circuit
    1. High
      a. Internal circuitry: magnetron
      b. External circuitry: Stabilitron
    2. Low Q: voltage-tunable magnetron (VTM)
  B. Traveling-wave circuit
    1. Forward wave: magnetron amplifier, or TPOM
    2. Backward wave
      a. Injection gun, nonreentrant beam: power BWO (M-BWO, M-type carcinotron); power backward-wave amplifier (M-BWA)
      b. Continuous cathode reentrant beam: Amplitron

* After James Holohan, *Aviation Age*, April, 1958.
† Not commercially available.

resistance to shock and vibration; and economic factors such as cost and availability. Because of the rapidly changing technical development of microwave tubes, it is not possible to catalogue these considerations at present, and this discussion is limited to the general features of the more important types. A few of the principal types of microwave tubes will be discussed.*

23. Microwave Triodes. Microwave triodes operate satisfactorily at frequencies up to about 4,000 Mc, at which the available output power drops to a few milliwatts

* See also Chap. 16, Electronic Oscillators.

Table 3. Microwave Tube Characteristics*

Tube type	Operating freq, kMc	Band width	Tunability	Power: Pulse	Power: C-w	Stability[1]	Noise[2]	Size and weight	Gain	Efficiency	Main uses	Outstanding feature
Microwave triodes and tetrodes	0.2–4	≅1%	Mechanical	mw to Mw @ u.h.f., mw to kw @ higher freq.	mw to > 10 kw @ u.h.f., mw to few watts above 3 kMc	Good	6 db @ 500 Mc narrow band, 10 db @ 1,000 Mc, 20 db @ 2,000 Mc	Low to high	Low	High at 1 kMc, low above 1 kMc	U-h-f transmitters and receivers; microwave relays	Simple, cheap
Klystron amplifiers	0.2–30	1.5–10%, 15% @ l.f.	Mechanical, 1%[4] or 15%[3]	100 kw (av), 10 Mw (p), 30 Mw @ S-band[D]	Several kw @ 3 kMc	Excellent	Medium	High	15–60 db, 90 db[D]	15–40%,[4] 55%[4,D]	Radar and communications transmitters	High gain, power amplifier, highest power capabilities
Reflex klystrons	1–70	.....	Electronic, 1%; mechanical, up to 3:1[3]	....	10 mw to 10 watts	Good	Medium	Very low	....	≅1%	Loc. osc., microwave relay transmitters, signal generators	Stable frequency source, simple, cheap, voltage-tunable
O-type forward-wave amplifiers (traveling wave tubes)	0.3–30	10%,[4] >2:1[3]	....	Up to 10 kw @ 3 kMc; 1 Mw @ 3 kMc[D]	mw to 100 watts @ 3 kMc	Good	6 db,[5] 3 db[5,D]	High with solenoid, moderate with magnet	50 db,[3] 30 db[2]	Up to 25%	Broad-band r-f power amplifiers, low-noise r-f amplifier for receivers and test equipment, microwave relays, ECM	Broad band width, low noise
O-type backward-wave oscillators (O-carcinotrons)	0.3–30, 200[D]	.....	Electronic, up to 3:1	....	mw to 1 watt, 100 watts @ 3 kMc[D]	Good	70 db	Moderate with magnet	....	1%	Voltage-tunable oscillators in receiver loc. osc., transmitter driver chains, signal generator	Quick, wide tunability
O-type backward-wave amplifiers	0.3–30	0.1 to a few %	Electronic, 2:1	....	mw	Good	6–10 db, 3.8 db @ S-band[D]	Moderate with magnet	15 db	<5%	ECM, tunable radar receivers	Narrow band width but voltage-tunable
Magnetrons	0.5–100	.....	Mechanical, 15%	mw to 5 Mw @ 3 kMc	mw to 10 kw @ kMc	Poor	High	Low	....	<70%	Radar and ECM transmitters	Efficient, high power, compact
Voltage-tuned magnetron	0.3–6	.....	Electronic, >2:1	....	Up to 1 watt	.....	Fair	Very low	....	<25%	Not yet in use; probable uses: loc. osc., low-power source for transmitter driver chains, signal generator	Simple voltage-tunable generator, very wide range, cheap, small

Tube type	Operating freq, kMc	Band width	Tunability	Power		Stability[1]	Noise[2]	Size and weight	Gain	Efficiency	Main uses	Outstanding feature
				Pulse	C-w							
M-type forward-wave amplifiers (TPOM, magneton amplifier)	0.3–37	25%	....	Up to 8 Mw @ 1.2 kMc	6 kw @ 1.2 kMc[D]	Good	Fairly high	Moderate with magnet	15 db	<50%	Pulse radar transmitter output	Efficient, broad band (possibilities of low noise, higher gain)
M-type backward-wave amplifiers	0.3–30[6]	Few %[6]	Electronic, 1.5:1[5]	Up to 150 kw @ 3 kMc[6]	10 mw to 1 kw @ 3 kMc[6]	....	....	Moderate with magnet	10 db	High[5]	....	High power and efficiency[5]
M-type backward-wave oscillators (M-carcinotrons)	0.3–15 mm range[D]	.....	Electronic, 1.5:1	Up to 150 kw @ 3 kMc	10 mw to 1 kw @ 3 kMc	Good	>25 db	Moderate with magnet	....	<50%	ECM and scatter transmitters	High power, voltage-tunable oscillator
Stabilitron	0.2–10	.....	Mechanical, 10%	200 kw to 8 Mw @ 1.5 kMc	Watts to 20 kw	Good	Fairly high	Moderate with magnet	....	<60%	Radar transmitters	Efficient, high power, stable frequency source
Amplitron	0.2–10	10%	....	200 kw to 8 Mw @ 1.5 kMc	Watts to 20 kw	Good	Fairly high	Moderate with magnet	10 db	<70%	High-level r-f amplifier for radar transmitters	Efficient, high power, broad band

* After James Holohan, *Aviation Age*, April, 1958.
[1] Frequency stability, *i.e.*, cps fluctuation per unit band width. In case of klystron, "excellent" refers to stability of tube as oscillator.
[2] Noise figure for amplifiers; signal-to-noise ratio for oscillators.
[3] Low-power types.
[4] Megawatt-power types.
[5] C-w types.
[6] Expected performance.
[D] In development.
NOTES: (I) The parameters listed in this table for a specific tube type obviously cannot all be gotten simultaneously with one tube. The figures represent the best or near-best that has been attained so far with tubes in a particular class. (II) Noise figures for "wave"-type amplifiers are for low-noise tubes. These are all limited in power-handling ability. (III) "Wave" tubes have the remarkable property of being relatively insensitive to load changes.

for c-w operation, or a few watts peak for pulsed operation. The chief limitation in modern microwave triodes is the long transit time of electrons in the grid-cathode space. The engineering approach to overcome this has been to design tubes with close interelectrode spacing and with a high equivalent grid-plate voltage. The factors which determine the ultimate limitation are the maximum available cathode current density, the small physical size of the structure required, and the maximum allowable tube-element temperature.

Fig. 18. Microwave triode structure.

Recent improvements in the high-frequency performance of triodes have resulted from an improvement in materials and techniques. For example, Fig. 18 shows the design of an experimental tube capable of operation up to 10 kilomegacycles. The use of ceramics and high-temperature getters permits this tube to be operated at high bulb temperatures so that the dissipation ratings can be increased. Further improvements in triodes can be expected when the current-density capabilities of cathodes are increased.

Because of the high operating efficiency and low noise characteristics of triodes at the lower end of the microwave band and because their cost is low compared with other tubes, the microwave triode is a particularly important member of the microwave-tube family.

24. Klystron Tubes. In contrast with microwave triodes, klystron tubes are dependent on transit-time effects for efficient operation. As discussed previously, a cavity resonator is used so that a small amount of drive power will cause a high a-c voltage to appear across the cavity gap. The resulting velocity modulation and bunching action produce a large a-c current in the beam. A second tuned cavity (Fig. 19*b*) is located a suitable distance from the first and a high voltage is induced across the catcher grids. Useful power is coupled out of the cavity by means of some coupling device, such as the loop illustrated, and is delivered to a load terminating the transmission line.

Increased amplifier efficiency results when one or more tuned cavities are placed between the input and output cavities. Figure 6 shows a three-cavity cascade klystron. The highest plate efficiency of these tubes is of the order of 30 to 45 per cent. Most klystrons are ½ to 20 per cent efficient. Their chief advantages are that the structures are able to dissipate large amounts of heat, and the cavities provide good frequency stability. Output powers range from a few milliwatts for low-voltage amplifiers to as high as 30 megawatts peak pulse power.[31]

Since the klystron bunching action produces large amplitude harmonics, if the second cavity is tuned to a higher harmonic, the klystron can be used as a frequency multiplier. Clearly, of course, the beam must be redesigned to ensure adequate transmission of the beam through the smaller cavity. Multiplication ratios of 10 to 20 are typically used.

If, as illustrated in Fig. 19*b*, a reflector turns the beam back upon itself, the cavity resonator acts as both the buncher and the catcher. The phase of the returning beam can be adjusted so that energy is extracted from the beam and delivered to the a-c field to produce oscillations. Small changes in the reflector voltage will result in a change in the phase between the a-c current in the beam and the a-c gap voltage. Thus, with some sacrifice in power, the oscillation frequency can be shifted slightly. A typical power and frequency characteristic of a reflex klystron is shown in Fig. 20.

Although the efficiency of the reflex klystron is of the order of only 0.1 to 10 per cent, its simple structure and ease of mechanical and electronic tuning make it a very useful signal source.

25. Traveling-wave Amplifiers. The traveling-wave amplifier is similar in many respects to the klystron. The velocity modulation occurring in the klystrons also occurs in the traveling-wave tube but in a more gradual manner since the impedance of the circuit is low. However, the interaction occurs over a longer distance so that the net effect is about the same in the two tubes. The electrons traveling down the tube axis are grouped so that the a-c field of the circuit slows down the bunched electrons, thereby transforming some of the kinetic energy of the electrons into useful r-f energy. The electron velocity should be slightly greater than the phase velocity of the wave propagation down the slow-wave structure. Fortunately, for many slow-wave structures the phase-velocity variation as a function of frequency is very slight; therefore high gain is possible over a frequency range of about an octave or more without any adjustment of the beam voltage or circuit.

Output powers of the order of milliwatts to kilowatts are attainable for continuous operation, megawatts are attainable for pulse operation, and noise figures as low as 6.5 db have been attainable over a relatively broad frequency range, which makes them useful as r-f amplifiers for receivers. New developments in traveling-wave tubes include light-weight tubes that are focused electrostatically or by means of periodic magnets.

(*a*)

(*b*)

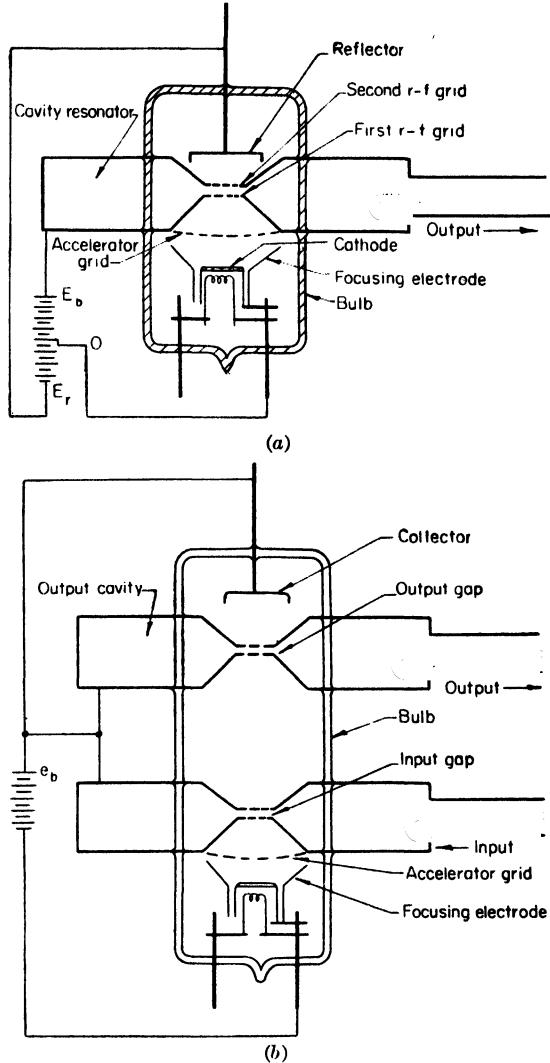



FIG. 19. (*a*) Schematic drawing of reflex klystron in glass bulb with resonant cavity, loop, and line attached. The space between the first and second r-f grids is the ***interaction gap***. Electrons become bunched in the ***drift space*** located between the interaction gap and the reflector. (*b*) Schematic drawing of two-cavity klystron. Bunching action is begun in the input gap. In the output gap the bunching of the electrons reaches an optimum value. This type of tube may be used for amplification at microwave frequencies or for frequency multiplication.

The traveling-wave tube is one of the most important amplifiers of the microwave-tube family because of its very wide band width, low noise, high power, and high gain. Its chief drawback is its relatively low efficiency. At high power levels, 20 to 25 per cent efficiencies are attainable, and at low power levels the efficiency drops to a few per cent

26. Backward-wave Tube.[32] The backward-wave tube is a device which can operate as a voltage-tunable oscillator or amplifier over a wide frequency range.

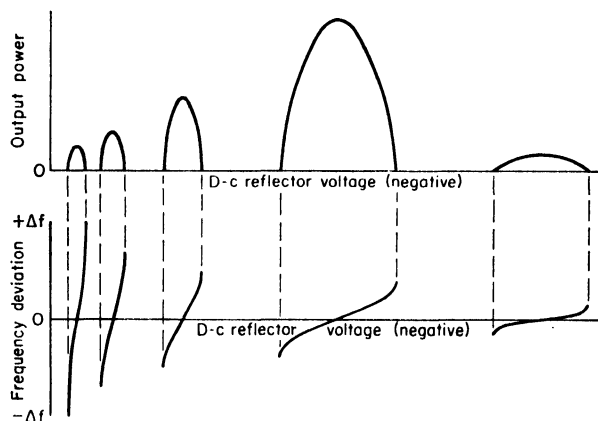



FIG. 20. Power output and electronic tuning of a reflex klystron.

Satisfactory operation over as high as a 3:1 frequency range and continuous output powers ranging from milliwatts to kilowatts have been obtained.

A periodic-type interaction structure, such as the helix, is used, but the circuit is designed to operate in the highly dispersive region where the group velocity is opposite to the phase velocity. For proper operation the phase velocity as a function of frequency should vary rapidly with frequency in the manner indicated in Fig. 14*b*.

A schematic diagram of the backward-wave tube is shown in Fig. 21. The r-f field of the backward-wave circuit has a phase velocity slightly less than the beam, and energy is transferred from the beam to the field in a manner similar to that occurring in the traveling-wave amplifier. The r-f energy flow on the circuit in this case is in a direction opposite to the phase velocity and is fed back to the input. For positive feedback the total phase shift around the loop must be an integral number of cycles. In terms of the phase velocity v_p, the average velocity of the electrons in the beam v_0, the angular frequency ω, and the length of the tube L, the condition for phase reinforcement is

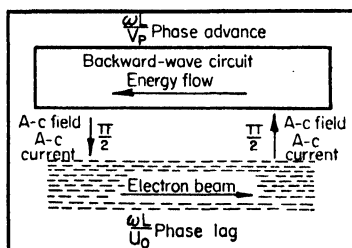



FIG. 21. Schematic circuit of a backward-wave tube.

$$\omega = \frac{(2n + 1)\pi}{[(1/v_p) - (1/v_0)]L}$$

where n is usually equal to zero. Since v_p depends on ω it is clear that ω is dependent on v_0. Since v_0, in turn, depends on the beam voltage, the device is therefore voltage-tunable.

If a signal is fed into the circuit at the right-hand end of the tube and if the tube is voltage-tuned to the frequency of the signal, the signal appearing at the left end of the tube will be amplified. For operation as an amplifier there must, of course, be enough

feedback to produce a net gain greater than unity but not so much as to produce oscillations. The effective Q of this amplifier is exceptionally high, permitting operation in the millimeter-wavelength bands. If the beam current is made relatively large, the loop gain becomes infinite and the tube becomes a voltage-tunable oscillator.

In addition to the important advantage of voltage tuning, the backward-wave oscillator, since it has no resonant circuit, is entirely immune from frequency pulling by the circuit. Frequency modulation can be obtained by varying the beam voltage and amplitude modulation can be obtained by varying the beam current.

27. Microwave Magnetron Oscillator.[13,23,24,28,29,33] The magnetron is one of the most important members of the microwave-tube family. Magnetrons are used extensively as sources of c-w power and high-pulse power. For pulse operation they are capable of delivering megawatts of peak powers in the lower-frequency ranges and milliwatts of power in the millimeter range. Pulse operation is accomplished by anode modulation with rectangular pulses of 0.5 to 10 μsec duration. For c-w operation, kilowatt power levels are readily obtainable at lower portions of the microwave frequency range. Load efficiencies as high as 70 per cent have been obtained.

FIG. 22. Paths of electrons in simple magnetron; *a*—electrons giving energy to the r-f field, *b*—electrons in absence of r-f field, *c*—electrons extracting energy from the r-f field.

The circuit of the microwave magnetron is similar to that illustrated in Fig. 13*b*. Two waves exist traveling in opposite directions. If a positive voltage is applied to the anode and a magnetic field normal to the electric field is applied, the electrons will rotate around the cathode with an average velocity approximately equal to the

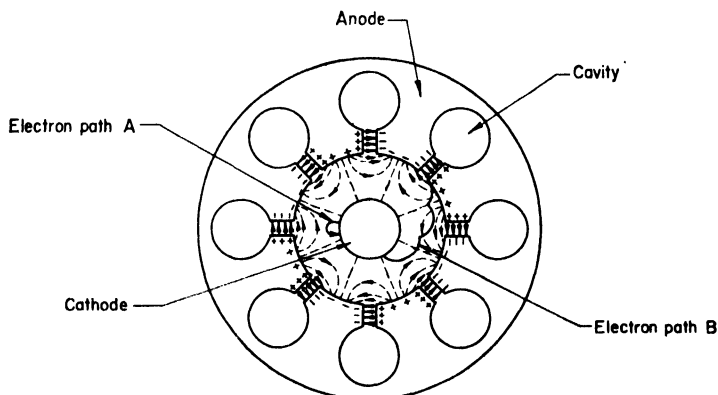



FIG. 23. Paths followed by unfavorable and favorable electrons in a magnetron.

ratio of the electric and magnetic fields. If this velocity is slightly greater than the phase velocity of one of the two waves, interaction will occur in a manner similar to that for the traveling-wave tube. An important difference, as illustrated in Fig. 22, exists for the magnetron which results in high electronic efficiency. Electrons which tend to absorb energy from the r-f field, and therefore are unfavorable electrons, are quickly returned to the cathode. On the other hand, the favorable electrons which contribute potential energy to the r-f field remain in the interaction space for a rela-

tively long time, finally striking the anode but with low kinetic energy. This electron-sorting mechanism results in the very high efficiency of the magnetron.

Figure 23 shows the path followed by an unfavorable (A) and a favorable (B) electron in an actual oscillating magnetron. Since the unfavorable electrons are returned, the resultant space charge appears as "spokes of a wheel" rotating at approximately the phase velocity of the circuit wave, as illustrated in Fig. 24. Since the magnetron operates using a highly resonant circuit, a reactive load will pull the frequency of oscillation.

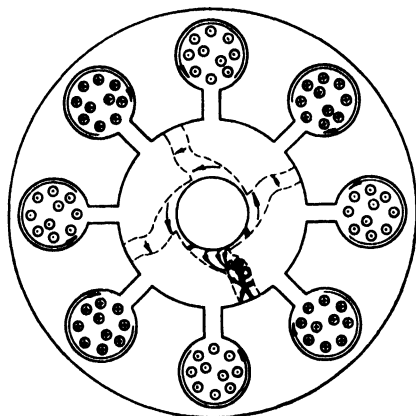



Fig. 24. Rotating space charge in magnetron.

While the magnetron is the most widely used microwave tube, free-running oscillators are losing popularity in many microwave systems. Short life has been a problem on some types, and the narrow range of mechanical tuning is a drawback. Considerable effort is being exerted to overcome these deficiencies. Notable among the new developments is the voltage-tunable magnetron (VTM) having characteristics highly competitive with the backward oscillator.[33]

28. M-type Tubes.[34] A schematic diagram of the M-type tube is shown in Fig. 11. In many respects this tube is similar to the traveling-wave, backward-wave, and magnetron tubes discussed. As with the two former types a slow-wave interaction structure is used, having the group and phase velocities in the same or opposite direction depending on the type of operation desired. A crossed E and H field is employed, as in the magnetron, to remove unfavorable electrons from strong interaction regions. An important additional advantage of this structure is that the cathode is separate from the interaction space and is not subject to the deleterious effects of back bombardment of electrons. As with the magnetron, the paths of the electrons are similar to those shown in Fig. 22. The kinetic energies of the three groups of electrons are approximately equal. The electrons following path a are contributing energy to the r-f field, but as for the magnetron this is a conversion of d-c potential energy into r-f energy. This energy exchange is in contrast with the klystron and traveling-wave tubes where the d-c kinetic energy of the electrons is converted into r-f energy.

Tubes operating as forward-wave amplifiers are often referred to as M-FWA, or as TPOM (*tube à propagation d'onde*). Efficiencies of 50 per cent are obtainable for tubes operating with band widths of 25 per cent and gains of approximately 30 db. New developments in this field include a coaxial-type structure, the center conductor of which carries a direct current to provide concentric lines of magnetic flux.

The backward-wave oscillators are referred to as M-BWO or as M-carcinotrons. These tubes have high efficiency, high power, and are voltage tunable. Their chief limitation, at present, appears to be noise. Very little work seems to have been done on the backward-wave amplifier (M-BWA).

29. New Microwave Tubes. The types of microwave tubes discussed so far have been applied very successfully in a wide variety of microwave systems. There is, of course, a constant demand for improved tube performance which results in an increasing number of new basic types of tubes. The most notable of the newer tubes in recent years have been of the crossed-field type, a few of which are discussed below.

Crossed-field types (also called M-types*) are tubes in which the electron beam flows normal to mutually perpendicular d-c magnetic and electric fields. The magnetron is an older member of this group. The newer types which are discussed include the

* See Glossary of Important Microwave-tube Terminology on page 8-26.

M-type forward-wave amplifiers, backward-wave amplifiers, and oscillators, the amplitron, and the Stabilitron.

30. Platinotron.[38] The Platinotron* is a crossed-field tube which can be used as an amplifier or as an oscillator. The tube, shown schematically in Fig. 25, is very similar to the magnetron but has two external connections, one for the input and one for the output. As shown in Fig. 25, the "spokes" of the beam travel in a direction opposite to the wave. The tube is essentially a backward-wave amplifier with electronic regenerative feedback.

FIG. 25. Mode of operation of a Platinotron is complex, utilizing crossed electric and magnetic fields and a reentrant slipping-stream electron beam to give backward-wave behavior. (*Electronics, August,* 1957, *p.* 164.)

When used as an amplifier the tube is referred to as an Amplitron.* In this case, power is fed to the input, and amplifier power is obtained from the output terminal. Efficiencies of 50 to 75 per cent are obtainable at approximately 10-db gain and 10 per cent band width. It is possible to operate the tube with pulse widths up to 20 μsec in duration.

When the tube is used as a self-excited stabilized oscillator, it is called the Stabilitron. Figure 26 is a schematic circuit of the device. Self-excitation results from the feedback from the mismatch in the output circuit and the reflection from the resonant circuit. The wave reflected to the left reaches the resonant circuit with no appreciable change in amplitude but with some phase shift. The wave reflected to the right is amplified by the tube and also experiences phase shift. By proper adjustment of the reflection coefficients and the phase shift around the loop, the tube can be made to oscillate. The use of a high-Q circuit allows for considerable stabilization. Tuning is accomplished by changing the resonant frequency of the stabilizing cavity. Over-all efficiencies of 45 to 60 per cent are obtainable.

FIG. 26. The Stabilitron.

* Trademark of Raytheon Manufacturing Company.

Glossary of Important Microwave-tube Terminology.[12, 23, 35, 36] The terms of the previous sections apply in addition to the following:

Admittance. The ratio of the a-c current flowing into the input terminals to the a-c voltage appearing across the input terminals. The circuit input admittance is the portion of the input admittance due to the circuit alone, and the electronic input admittance is the portion of the input admittance due to the electron beam alone.

Transadmittance. The ratio of the a.c. flowing from the tube output terminals to the a-c voltage appearing across the input terminals.

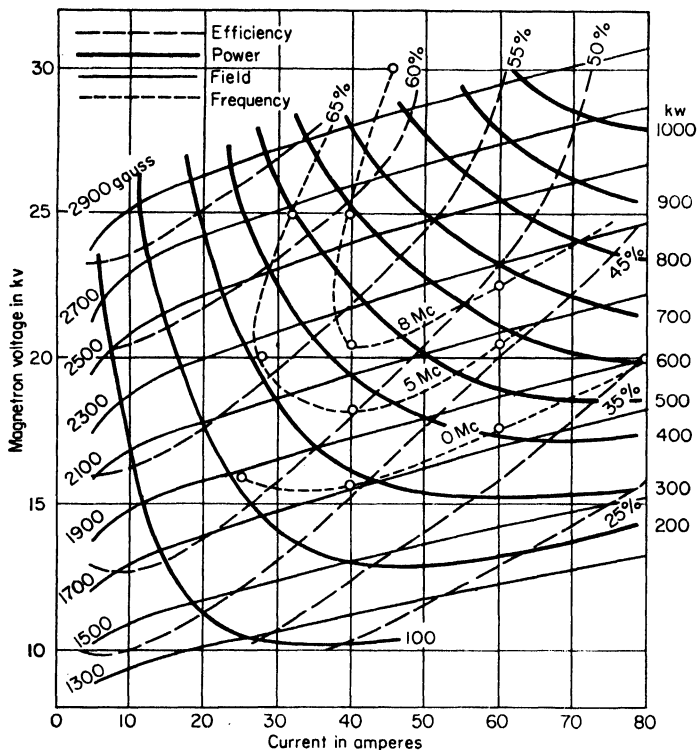



FIG. 27. Performance chart (of a magnetron).

Pulsed Frequency Spectrum and Band Width. The frequency spectrum of a pulsed oscillator is a plot of the r-f energy distribution vs. the frequency of oscillation.

The frequency-spectrum band width is the frequency interval between two points of equal power (usually the quarter-power points) on the frequency spectrum.

Efficiency. Electronic efficiency. The ratio of the total r-f power generated to the d-c power in the electron beam.

Circuit efficiency. The ratio of the r-f power delivered to the load to the total r-f power generated.

Load efficiency. The ratio of the r-f power delivered to the load to the d-c power in the electron beam. Also, it is the product of electronic efficiency and circuit efficiency.

Frequency Pulling and Pushing. Frequency pulling is the change in frequency of oscillation caused by a change in load admittance. The *pulling figure* of an oscillator is the maximum change in frequency which occurs when the phase angle of the load

(usually adjusted to have a VSWR of 1.5) is varied through at least one-half of a guided wavelength.

Frequency pushing is the change in frequency of oscillation due to a change in anode current. The *pushing figure* of an oscillator is the maximum change in frequency which occurs when the anode current is changed a certain amount (not including frequency changes due to thermal effects).

Gain. The ratio of the output power to the input power. For small input signals, the gain is referred to as the small-signal gain. At large signals the gain is often compressed and the gain is referred to as the large-signal gain.

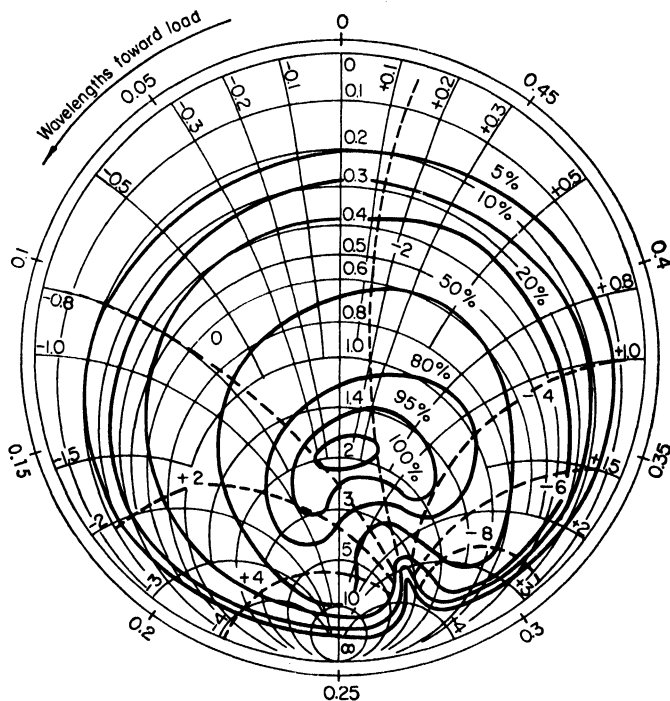



FIG. 28. Rieke diagram (of a reflex oscillator).

M-type Tubes. Wave-type tubes in which the electron beam flows normal to mutually perpendicular d-c magnetic fields. These tubes are also called *crossed-field tubes.*

Noise and Spurious Responses. The noise figure is the ratio of the available signal-to-noise ratio at the amplifier input terminals to that available at the output terminals.

The signal-to-noise ratio of an oscillator is the ratio of the r-f power generated at the oscillator frequency to the noise power contained within a specified frequency band or bands.

Spurious responses are miscellaneous unwanted signals appearing in microwave amplifiers and oscillators because of various causes such as ion phenomena in the beam, coupling between undesired circuit harmonics and beam harmonics, microphonics, heater hum, and leakage.

O-type Tube. Wave-type tubes in which the electron beam flows parallel to the magnetic field (O-type refers to ordinary-type tubes).

Performance Chart. A plot of anode voltage vs. anode current for constant values of magnetic field, constant output power, and constant load efficiency (see Fig. 27).

Q. The Q of a resonant circuit for a microwave tube is defined as 2π times the ratio of the electromagnetic energy stored per cycle to the energy lost per cycle at the resonant frequency.

The *loaded Q* of a resonant circuit is 2π times the ratio of the electromagnetic energy stored per cycle to the energy lost per cycle in the internal losses and the coupled-in external load. The *unloaded Q* is the value one would obtain if the energy were lost in the internal losses only. The *external Q* is the value one would obtain if the energy were lost in the coupled-in external load only. These definitions give rise to the relationship

$$\frac{1}{Q_{\text{loaded}}} = \frac{1}{Q_{\text{unloaded}}} + \frac{1}{Q_{\text{external}}}$$

Rieke Diagram. A chart showing the loci of load admittances giving constant power and constant frequency. The chart is often plotted on a Smith chart (see Fig. 28).

Tuning. The *tuning range* is the frequency interval over which the tube is considered operable. The *mechanical tuning range* is the frequency interval over which it can be mechanically tuned, and the *thermal tuning range* and *electronic tuning range* are the intervals over which the tube can be tuned by thermal or electronic means.

References

1. Danforth, W. E.: "Thorium Oxide and Electronics," in "Advances in Electronics," edited by L. Marton, vol. 5, pp. 169–211, Academic Press, Inc., 1953.
2. Beck, A. H., A. D. Brisbane, A. B. Cutting, and G. King: New Type of Diffusion Cathode, *Elect. Commun.*, 172, September, 1955.
3. Lemmens, H. J., M. J. Jansen, and R. Loosjes: New Thermionic Cathode for Heavy Loads, *Philips Tech. Rev.*, **11**, 341, June, 1950.
4. Esperson, G. A.: L-Cathode Structures, *Proc. IRE*, **40**, 284, March, 1952.
5. Levi, R.: "New Dispenser Type Thermionic Cathode," Convention Record of the IRE 1953 National Convention, Part 6, Electron Devices.
6. MacNair, D., R. T. Lynch, and H. B. Hannay: Molded Thermionic Cathodes, *J. Applied Phys.*, **24**, 1335, October, 1953.
7. Pierce, J. R.: "Theory and Design of Electron Beams," Van Nostrand, 1949.
8. Pierce, J. R., and W. G. Shepherd: Reflex Oscillators, *Bell System Tech. J.*, **26**, 460, July, 1947.
9. Müller, M.: New Points of View in the Design of Electron Guns for Cylindrical Beams of High Space Charge, *J. Brit. IRE*, 1958, p. 53.
10. Danielson, W. E., J. L. Rosenfield, and J. A. Saloom: Analysis of Beam Formation with Electron Guns of the Pierce Type, *Bell System Tech. J.*, **35**, 375, March, 1956.
11. Heil, O., and J. J. Ebers: A New Wide-range, High-frequency Oscillator, *Proc. IRE*, **38**, 645, June, 1950.
12. Hamilton, D. R., J. K. Knipp, and J. B. H. Kuper: "Klystrons and Microwave Triodes," Vol. 7, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1948.
13. Spangenberg, K. R.: "Vacuum Tubes," McGraw-Hill, 1948.
14. Learned, V., and C. Veronda: Recent Developments in High-power Klystron Amplifiers, *Proc. IRE*, **40**, 465, April, 1952.
15. Wang, C. C.: Electron Beams in Axially Symmetrical Fields, *Proc. IRE*, **38**, 135, February, 1950.
16. Wang, C. C., and K. Bol: Sperry Gyroscope, unpublished (O.N.R. sponsored research).
17. Pierce, J. R.: Some Recent Advances in Microwave Tubes, *Proc. IRE*, **42**, 1735, December, 1954.
18. Watkins, D. A.: Traveling-wave Tubes and Backward-wave Tubes, *Proc. of the Symposium on Modern Advances in Microwave Techniques, Polytechnic Institute of Brooklyn*, **IV**, 83, November, 1954.
19. Peter, R. W.: Problems of Producing Low-noise Electron Beams, *Proc. of the Symposium on Modern Advances in Microwave Techniques, Polytechnic Institute of Brooklyn*, **IV**, 159, November, 1954.
20. Pierce, J. R., and W. E. Danielson: Minimum Noise Figure of Traveling-wave Tubes with Uniform Helices, *J. Applied Phys.*, **25**, 1163, September, 1954.
21. Moreno, T.: "Microwave Transmission Design Data," McGraw-Hill, 1948.
22. Marcuvitz, N.: "Waveguide Handbook," Vol. 20, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1951.
23. Collins, G. B.: "Microwave Magnetrons," Vol. 6, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1948.
24. Boulding, R. S. H.: "The Resonant Cavity Magnetron," Van Nostrand, 1952.
25. Pierce, J. R.: "Traveling Wave Tubes," Van Nostrand, 1950.
26. Llewellyn, F. B., and L. C. Peterson: Vacuum-tube Networks, *Proc. IRE*, **32**, 144, February, 1944.

27. Llewellyn, F. B.: "Electron Inertia Effects," Cambridge University Press, 1941.
28. Beck, A. H. W.: "Thermionic Valves," Cambridge, 1953.
29. Reich, Ordung, Krauss, and Skalink: "Microwave Theory and Techniques," Van Nostrand.
30. Wang, C. C.: Linear Beam Tube Theory, *IRE Trans. Electron Devices*, **ED-4** (1), 92, January, 1957.
31. Chodorow, M., E. L. Ginzton, I. R. Neilsen, and S. Sonkin: Design and Performance of a High-power Pulsed Klystron, *Proc. IRE*, **41,** November, 1953.
32. Heffner, H.: Backward-wave Tube, *Electronics*, October, 1953.
33. Wilbur, D. A.: New Magnetron Principles, *Proc. of the Symposium on Modern Advances in Microwave Techniques, Polytechnic Institute of Brooklyn*, **IV,** 133–143, November, 1954.
34. Warnecke, R. R., P. Guenard, O. Doehler, and B. Epsztein: The "M"-type Carcinotron Tube, *Proc. IRE*, **43,** 413, April, 1955.
35. IRE Standards on Electron Devices: Microwave Terms (56 IRE 7.51), *Proc. IRE*, **44,** 346, March, 1956.
36. "Reference Data for Radio Engineers," 4th ed., Federal Telephone and Telegraph Corp., 1957.
37. Bibliography in Chap. 19, Sec. 159.
38. Brown, W. C.: Platinotron Increases Search Radar Range, *Electronics*, August, 1957.

CHAPTER 9

SEMICONDUCTOR DIODES

BY W. F. PALMER[1]

1. General. These two-element electron devices are available in a wide variety of forms. All depend upon a contact or junction between two suitable dissimilar metals

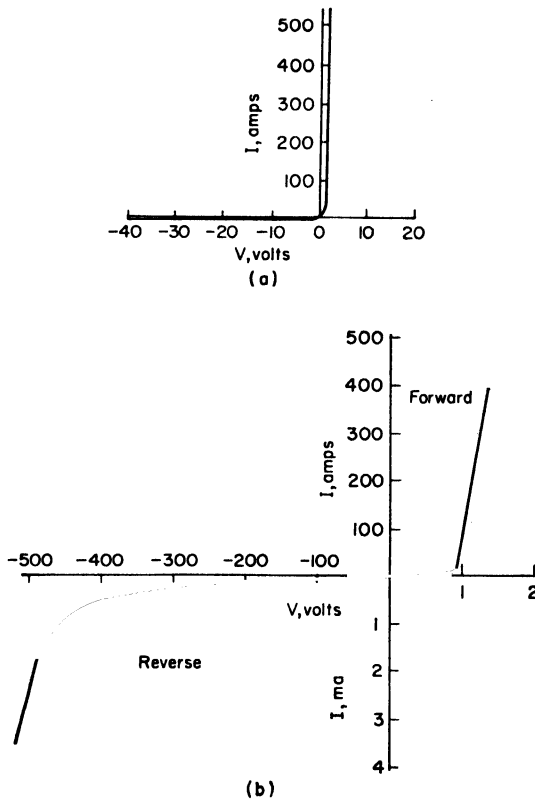



FIG. 1. Volt-ampere characteristics of a semiconductor power-rectifier diode. (*a*) Equal scales. (*b*) Expanded scale.

or certain other compounds (as metallic oxides). A potential barrier which appears at this junction or interface is responsible for the nonlinear or nonohmic volt-ampere conduction characteristic which results. It is this nonlinear characteristic which

[1] Sylvania Electric Products Inc.

makes possible the use of these devices as rectifiers or detectors. Figure 1 illustrates the nonohmic response of a semiconductor rectifier. Note the nearly ideal diode characteristic. Semiconductor diodes are made from a wide variety of materials or compounds and in various sizes ranging from subminiature types to large rugged power rectifiers. They may be used to detect minute r-f signals or to rectify kilowatts

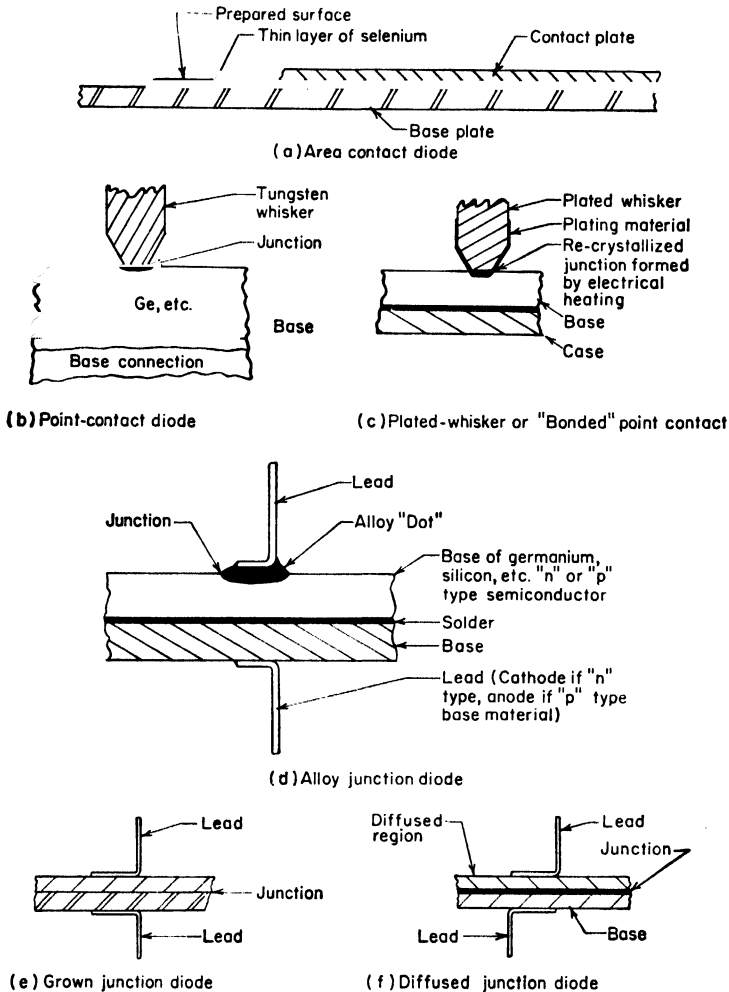



FIG. 2. Various semiconductor constructions.

of low-frequency a-c power. Figure 2 shows typical construction of several types of diodes.

Various classes of diodes and their major characteristics are given in Table 1.

2. Historical Note. The earliest common use of a semiconductor detector was that of crystalline lead sulfide (galena) in early radio reception. Later, copper-sulfide and copper-oxide devices were used as rectifiers and photocells. In the 1930's, research into semiconductor theory led to the development of metallic rectifiers, selenium being

the first of commercial importance, followed by silicon and germanium point-contact diodes in the 1940's and later by the large-area silicon and germanium types which are becoming commercially available at present.

Table 1. Various Types of Semiconductor Diodes

Construction	Material	Power	Frequency	Application notes
Point contact	1. Germanium	Low to medium	To 250 Mc, higher in appropriate designs	Detection, frequency conversion, etc.; clippers; clamps; rectifiers within ratings
	2. Plated whisker or "bonded"	Low to medium	To 250 Mc, higher in appropriate designs	Higher forward conduction than (1), applications similar to (1)
	3. Silicon	Low to medium	To 500 Mc	Applications as for (1) but may have exceptionally low recovery time
	4. Silicon microwave	Low to medium	To many 1,000 Mc in appropriate designs	Radar detectors, mixer, etc.
	5. Selenium*	Low to high	To 10 Mc or more in appropriate designs	Primarily as a rectifier
Junction	6. Germanium	Low to high	To 10 Mc or more in appropriate designs	As in (1) above, longer recovery time than point-contact types, power rectifiers
	7. Silicon	Low to high	To 10 Mc or more in appropriate designs	As for (6) above, sharp reverse breakdown or saturation voltage permits use as regulator

* Has been classified as a junction but may be considered as a large-area contact type.

3. Materials. Materials from which semiconductor diodes are commonly made are listed below. Many others of less commercial significance are not included:

1. Selenium—widely used in power rectifiers and photocells
2. Germanium—becoming commonly used in a wide range of diodes
3. Silicon—becoming commonly used in a wide range of diodes
4. Copper sulfide—not widely used at present
5. Intermetallic compounds—experimental only at present but potentially useful in future types
6. Cadmium sulfide—primarily used in highly sensitive photocells

DIODE APPLICATIONS

Applications of semiconductor diodes may be classified as detectors, rectifiers, switches, and logic elements in bistable circuits. They may also be used as variable impedances.

4. Detectors. These are diodes used for the detection of alternating signal currents—usually at relatively low power levels and relatively high frequencies. Detection is nonetheless rectification of an alternating current, but this term is usually reserved for applications in which information in one form or another is carried by the signal. Typical use of point-contact diodes in amplitude-modulation detectors is shown in Fig. 3.

5. Detection Efficiency. Detection efficiency varies considerably with the applied signal voltage, as does linearity. In general, both are improved as signal level is increased. At relatively high levels the forward current produces a negligible voltage drop across the diode, but this drop becomes significant at lower levels. At very low levels, the diode does not conduct to any great extent and the nonlinearity of its transfer characteristic is not pronounced, so that rectification efficiency is very low, and the diode is a poor detector.

Figure 4*a* shows the calibration curve of a typical germanium point-contact diode used as an r-f detector. In the region between zero and approximately 0.25 volt, the output voltage is proportional to the square of the input, but it becomes linear at higher input voltages.

FIG. 3. Typical a-m detectors.

Detection efficiency also varies with frequency but is nearly constant to several hundred megacycles for point-contact types (see Table 1). Figure 4*b* shows a typical frequency-response curve for a germanium point-contact diode. That of certain silicon types may be somewhat better.

With suitable attenuators and amplifiers, semiconductor diode detectors may be useful over a wide range of input voltages (1 mv or less upward) and frequencies (zero to thousands of megacycles).

FIG. 4. (*a*) Germanium-diode rectification curve. (*b*) Detector output vs. frequency.

6. Rectifiers. These are diodes which are used to convert a-c power to d-c power for operation of equipment which cannot be operated from the former. (For example, d.c. is usually required for operation of tube or transistor amplifiers, electrolytic refining of certain metals, d-c motors, etc.) Rectification of lower-frequency power at relatively high levels is most common. During one half cycle of the applied a-c voltage, polarity is such as to bias the diode "on," and current flows in the diode-load circuit. During the other half cycle the diode is biased "off" and essentially no current flows in the load. Where the ripple component may not be tolerated, rectification is commonly followed by suitable filters. The high rectification efficiency and lack of heated cathodes are important advantages over thermionic rectifiers.

7. Rectification Ratio. The rectification ratio of a diode is the ratio of its d-c reverse resistance to its d-c forward resistance at arbitrarily specified voltages. These voltages are usually chosen to be representative of peak or average voltages encountered in a particular application or class of applications. If these points are wisely chosen, the rectification ratio gives a fair indication of the rectification efficiency likely to be obtained in actual application. A high rectification ratio is desirable and indi-

FIG. 5. D-c restorer.

cates highly efficient rectification or detection at the particular current and voltage levels.

Rectification ratios at low voltages and high currents may not accurately indicate the actual efficiency of rectification (Sec. 5), however, as the diode forward voltage may be a significant fraction of the peak input voltage. Under these conditions, germanium or selenium diodes may be more efficient than silicon types even though the latter have high rectification ratios. Figure 6 illustrates this point.

8. Limiting of Peak Currents in Rectifiers. When rectifiers are operated with a capacitive load, surge and peak currents may be greater than the diode ratings. The insertion of an impedance in series with the diode will often permit use of the diode by increasing the conduction angle and lowering the peak and surge current to safe values without seriously reducing output rectified power. Where resistive losses may be excessive, a series inductor will be useful and essentially becomes the filter input choke.

Since calculation of peak and surge currents is tedious, requiring knowledge of a great many circuit parameters, an experimental method of testing and modifying the circuit during design is given in Fig. 7. Diode current is sensed by R_1 and displayed on the calibrated oscilloscope. Placing the resistor in the common or ground lead will permit grounding the oscilloscope. Where a differential oscilloscope is available, the resistor may be placed directly at the diode terminals, also permitting the oscilloscope to be grounded. R_1 may also function as the limiting resistor or as a separate low-resistance shunt.

Diode voltage may also be observed if the diode is placed near ground potential or if the oscilloscope is floated. (Observe suitable precautions if high voltages are involved.)

FIG. 6. Diode rectification with resistive load.

Instantaneous dissipation can also be observed if voltage and current are displayed on the different oscilloscope axes. Dissipation hyperbolas may be drawn on the display screen to facilitate reading. Note that the isolating network R_2d_2 is needed to remove inverse voltage across the diode when this is sufficient to damage the oscilloscope or affect its accuracy of voltage readings. R_2, of course, must be considered as an input attenuator

FIG. 7. Rectifier current and dissipation test circuits. (*a* and *b*) Oscilloscope grounded; (*c* and *e*) oscilloscope floating; (*d*) transformer secondary grounded.

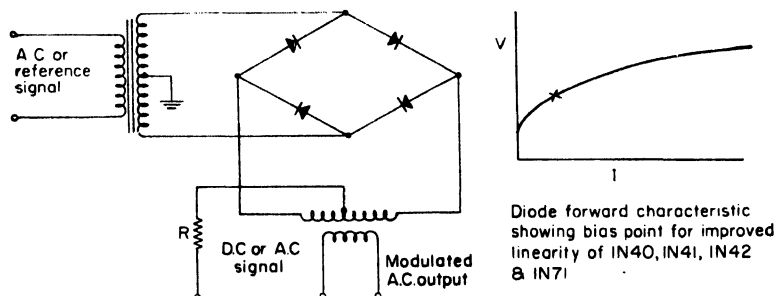



FIG. 8. Typical ring or bridge modulator.

to the oscilloscope, and calibration should be made with the network in place. Diode d_2 is a scale switch which protects the oscilloscope from diode inverse voltage without affecting its reading of forward rectifier voltage, and its "off" resistance should be much greater than R_2. A small-area silicon-junction type will be suitable.

9. Switches. Diodes used in a variety of nonrectifying applications may be termed switches, though it is perfectly correct to consider both rectifiers and detectors as switching applications of diodes.

When reverse-biased band conduction is very low, the diode is said to be in the "off" condition, and, conversely, when forward-biased into a high-conduction region it is said to be "on." Thus, the analogy with a mechanical switch is good. Unlike a mechanical switch, relay, vacuum tube, or transistor, no isolated input is available for control purposes, and the diode can be switched from state to state only by reversal of the potential across its terminals. This ability to become a near short circuit when forward-biased leads to the use of suitable diode types in clipping, clamping, chopping, d-c level restoring, and a host of similar applications. Figures 5 and 8 illustrate use

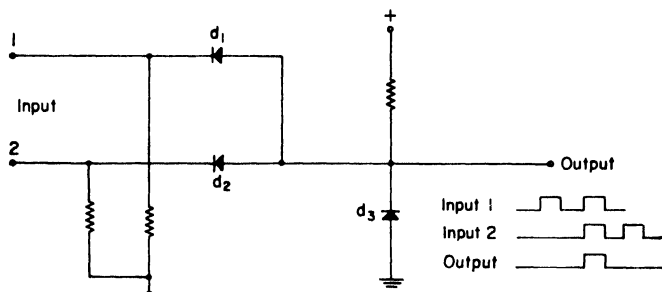



FIG. 9. Basic "and" gate.

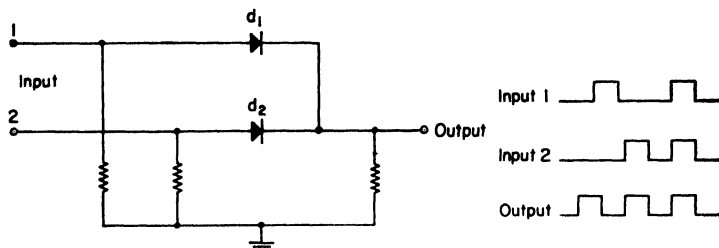



FIG. 10. Basic "or" circuit.

of a semiconductor diode in a d-c level restorer and a ring modulator or chopper, respectively.

Logic circuits using diodes may be considered another class of switching applications. Since diodes have two distinctly differing impedance levels, they lend themselves to use in performing predetermined decision or logic functions just as switches or relays may. Figures 9 and 10 show two basic logic functions which may be performed with diodes. Very complicated logic circuits involving many thousands of diodes have been assembled for use in electric computers. These examples indicate the ability of such circuits to make "logical decisions" based on "instructions" built into the particular circuit.

In the case of Fig. 9, all three diodes are normally biased "on" (conducting), d_3 serving to clamp the output to "zero"[1] or ground. If d_1 or d_2 is biased "off" by applying a positive or "one" pulse to one input, the other diode continues to conduct and the output remains at "zero." However, if both d_1 *and* d_2 are turned "off" by positive input signals the output also becomes positive, indicating a "one" stage. That is, a positive pulse or "one" appears at the output only if a one is applied to inputs 1 *and* 2.

[1] In Boolian, or binary, algebra the terms "one" or "1" and "zero" or "0" are used to indicate a "high" or "low," or "on" or "off" condition in a particular logic block or stage.

In the case of the particular "or" circuit of Fig. 10, both diodes are normally "off" and the output is a zero unless a positive pulse is applied to either input 1 *or* 2, in which case the diode conducts and a "one" appears at the output.

10. Choice of a Switching Diode. In general, a low forward voltage and low reverse leakage at the peak "on" current and peak "off" voltage are required.

A diode may be considered a good switch if:

1. The peak "on" voltage (at a specified "on" current) is much less than its peak "off" voltage.
2. Its reverse leakage (at the specified "off" voltage) is much less than the peak "on" current, *i.e.*, $V_{on} \ll V_{off}$ and $I_{off} \ll I_{on}$.

Germanium diodes are generally better than silicon as high-current low-voltage switches because of their low forward voltage, though silicon types are better for high-voltage low-current applications, as they have lower reverse or "off" leakage current, equivalent types assumed.

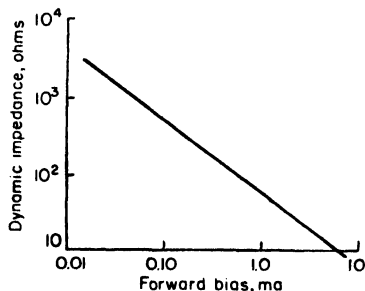



FIG. 11. Dynamic impedance characteristics, 1N527.

11. Variable-resistance Diodes. A forward-biased diode has an incremental series resistance which is inversely proportional to forward current or voltage. By varying the forward bias current by a suitable means, a diode may be used as a variable-resistance element in many applications. Figure 11 shows a graph of dynamic resistance as a function of current for one such diode (small-area type) which is controlled for this characteristic.

12. Variable-capacitance Diodes. The capacitance of a reverse-biased diode varies inversely with voltage (approximately as the inverse square root), and by varying this reverse voltage the diode shunt capacitance can be varied electrically. Thus such diodes can be used as "reactance tubes" in suitable circuits. Figures 12 and 13 show the voltage-capacitance curve for a large-area diode and a basic voltage-tunable circuit in which such a diode may be used.

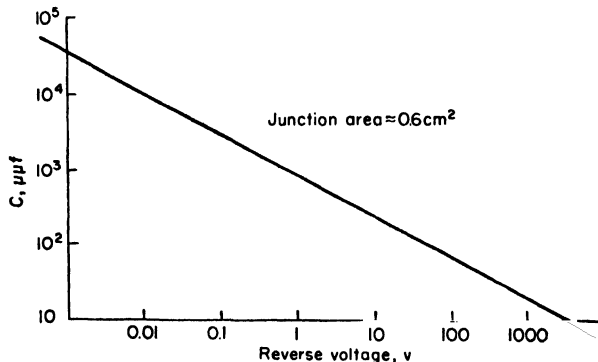



FIG. 12. Variation of capacitance of germanium junction diode vs. reverse bias voltage.

13. Frequency Multipliers or Harmonic Generators. Semiconductor diodes may be used as the nonlinear element in harmonic-generating circuits. For example, the output of a tube or transistor oscillator operating at moderately high frequencies is fed to a diode which feeds a circuit resonant at a harmonic of the oscillator frequency (*e.g.*, a v-h-f oscillator used with a diode to drive a u-h-f tank circuit).

In general, the output of such a circuit is relatively low when high multiplication factors are used but is usually high enough for low-level converter applications. Multiplication factors on the order of 3 to 5 are practical, though diode performance may be variable at higher values unless diode types having specified performance characteristics are used.

A typical frequency-multiplier-circuit arrangement is shown in Fig. 14.

14. Frequency Converters. Semiconductor diodes are used over a very wide frequency range (a-f to microwave) as frequency converters. Circuit considerations are similar to those involved in design of circuits for use of thermionic diodes. For linear conversion or detection, oscillator injection current should be much larger than the

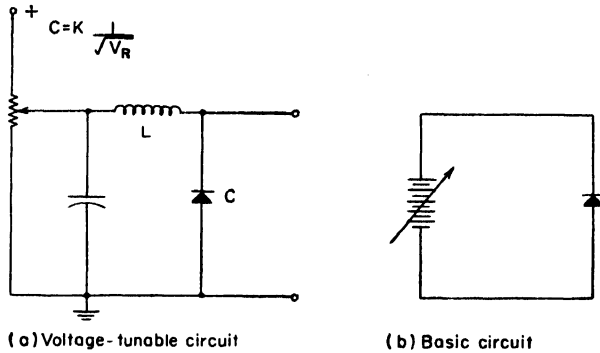



FIG. 13. Junction diode used as voltage-variable capacitor.

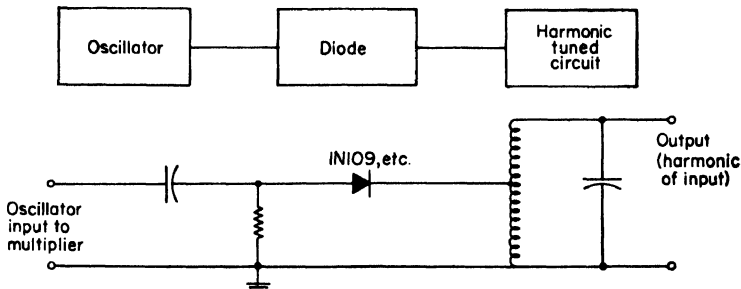



FIG. 14. Diode frequency multiplier.

signal current. At high frequencies the low capacitance of semiconductor diodes, especially of point-contact types, is desirable.

15. Semiconductor Diodes as Thermistors. Semiconductor diodes may be used as either low-impedance or high-impedance (current sources) thermistors by operating them in the forward or reverse directions, respectively. Forward-biased diodes have voltage saturating characteristics above their knee voltage and may be used as reference or regulating devices. However, this forward saturation voltage varies with temperature, the coefficient being approximately −2.5 mv per °C for germanium and −1.7 mv per °C for silicon. These values may vary somewhat with the design of the diode but indicate that diodes may be used as compensating elements where a negative voltage coefficient is desired.

Diode thermistors differ from the conventional type in that the characteristic is nonlinear, and if a-c signals are involved, they may be useful only at levels below their knee voltage. If the peak signal swing exceeds this value, distortion will result from nonlinear operation, or if operated on the plateau region, very little variation in dynamic resistance will be obtained even though the plateau voltage is shifting (and

the d-c resistance with it). At low levels, where a-c operation is practical, the voltage coefficients are variable with voltage and are lower than the values indicated above.

Thus, the diode thermistor is more useful in d-c compensation circuits. One specific application is the use of a diode in the base bias network of transistor circuits to improve temperature stabilization of collector bias current, as shown in Fig. 15 Shaping networks may be used as with thermistors to obtain the desired characteristic

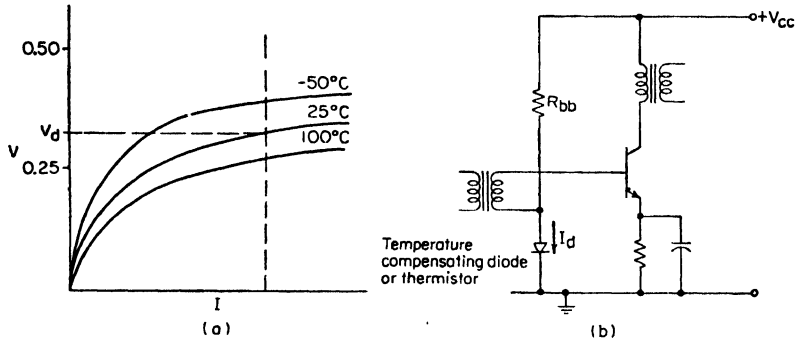



FIG. 15. Variation of forward voltage with temperature and use as thermistor, germanium diode.

Reverse-biased diodes operated in the reverse current saturation region where dynamic resistance is very high may be used as current sources which are temperature variable (Fig. 16). If a voltage which varies over a wide range is applied to such a diode and a series load, the current in the load is nearly constant but follows the temperature law of the particular diode used. Shaping networks may be used to adjust this characteristic as desired.

Diode reverse currents are also discussed in Sec. 28.

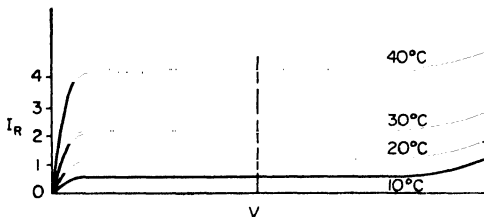



FIG. 16. Variation of reverse current with temperature, germanium diode.

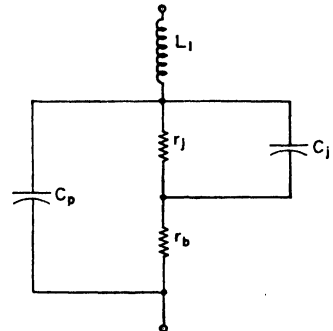



FIG. 17. Equivalent circuit of a semiconductor diode.

DIODE CHARACTERISTICS

16. Equivalent Circuit of a Semiconductor Diode. The equivalent electric circuit of a semiconductor diode is shown in Fig. 17. The actual values of the parameters depend upon the particular diode design, which varies considerably. It is important to note that the junction resistance and capacitance are highly variable with bias voltage (see Sec. 11).

At lower frequencies, lead inductance, package capacitance, and junction capacitance can often be neglected so that only the junction and base resistances need be considered for the forward-biased diode. At reverse biases (below breakdown), r_b is negligible compared with r_j, which is typically very large.

At higher frequencies it becomes necessary to consider the effect of other parameters,

and the utility of the diode may be limited by package and junction capacitance as well as lead inductances at several hundreds or thousands of megacycles.

Typical ranges of values of these parameters are (order of magnitude only):

L_l = lead inductance, near zero to several $\mu\mu$h
C_p = package capacitance, 0.5 to tens of $\mu\mu$f
r_b = lead and base resistance, 1 milliohm to 1,000 ohms
r_j = junction resistance (reverse), 10^3 to 10^{10} ohms
(forward), near zero to thousands of ohms
C_j = junction capacitance, 10^{-13} to 10^{-8} farad

17. Selenium Diodes. Selenium diodes are made by placing (usually by evaporation or flow coating) a thin layer of selenium on a backing plate which provides mechanical strength and serves as one electrode. The outer surface of the selenium layer is then treated to form a barrier layer. A pressure, or sprayed metal electrode, provides the second contact. A typical construction is shown in Fig. 2*a*.

Photocells of high sensitivity can be made by applying a transparent electrical contact to this prepared selenium surface. These diodes are normally used as photovoltaic cells but may occasionally be used as photoresistive devices. Selenium diodes

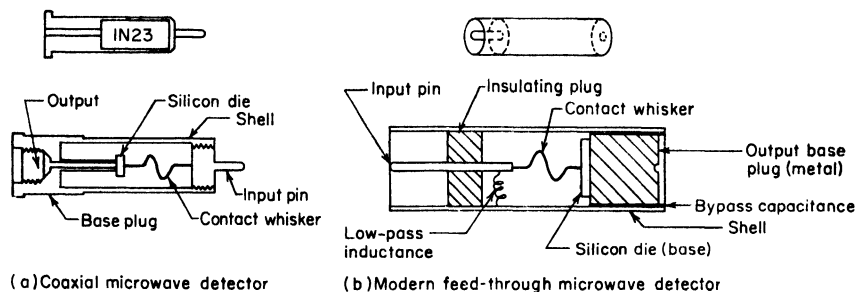



FIG. 18. Microwave detection diodes.

have relatively high reverse leakage and capacitance and are therefore most widely used in power-frequency rectification applications. Miniature designs, however, may be useful as high-frequency detectors and in other medium-speed applications in which germanium point-contact and small-area junction types are normally used. Forward conductance and rectification efficiency are much lower than those of equal-area germanium or silicon devices.

Failure of selenium rectifiers, when it occurs, is normally caused by a gradual reduction in forward conductance caused by reduction in contact area as elemental diodes become open-circuited, which results in increased forward dissipation (load current assumed constant). This in turn causes increased reverse leakage and dissipation leading to runaway and catastrophic failure ("burnout"). This is in contrast to the case of germanium- and silicon-junction diodes where increasing reverse leakage is the most significant cause of failure.

18. Silicon-junction Diodes. Silicon-junction diodes are available in alloy, grown, or diffused types of structures and in a wide range of sizes and ratings. They are of particular interest because of their very low reverse leakage and saturation currents, their abrupt reverse voltage breakdown, and the ability to withstand higher temperatures than those made of other common materials. A disadvantage in some low-voltage applications is the high "knee" voltage of the forward characteristic, approximately 0.7 volt vs. 0.3 volt for germanium.

Reverse leakage resistances are several orders of magnitude greater and reverse saturation current several orders of magnitude lower than those of germanium. Thus silicon diodes are highly desirable where very low reverse leakage is required (see Sec. 28).

The reverse breakdown voltage may also be as high as several thousand volts (per cell) in certain designs. The abrupt breakdown characteristics for which some designs are noted permits their use in various reference and regulating applications.

19. Silicon Reference or Regulating Diodes. Silicon-junction diodes may exhibit very abrupt reverse saturation or "breakdown" voltage characteristics. Such diodes may be used as sources of reference voltages or as shunt regulators in the same manner

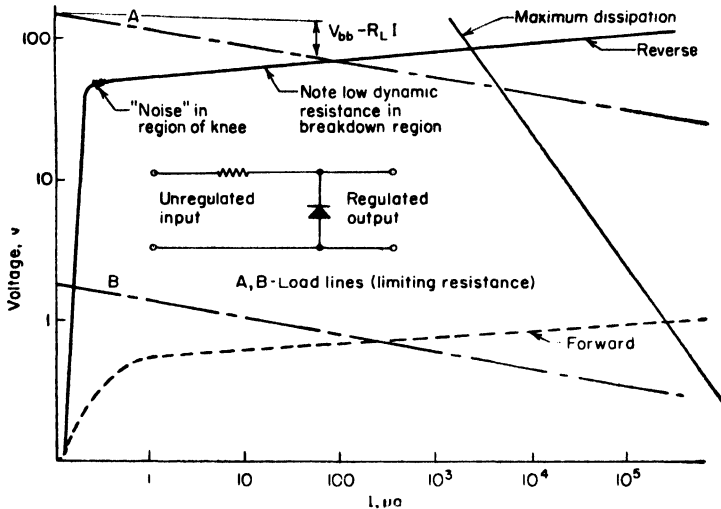



Fig. 19. Forward and reverse characteristics of a silicon reference diode.

as certain gas-filled tubes. The characteristic of a typical silicon reference diode is shown in Fig. 19. Note that the relatively high starting voltage needed for the glow-discharge tube is not required.

Diodes are available which have breakdown voltages from 1.5 to 600 or higher. Such diodes may also be biased in the forward direction to obtain voltages of about 0.75 to 1.5, as relatively little conduction occurs at lower voltages. (This is also true for diode types of other basic materials.) It is also possible to construct diodes having forward and reverse characteristics which are more or less symmetrical and which are useful in certain limiting applications (Fig. 20). If the characteristics are not symmetrical, a (reverse) bias voltage can be used to set the quiescent operating point midway between the forward and reverse conduction voltages. Two matched diodes can also be used if connected back to back. Thus these diodes may be used to regulate voltages or as reference voltage sources over a much greater range of voltage than gas tubes.

Fig. 20. Characteristics of diode having symmetrical forward and reverse voltage saturation characteristics.

20. Dynamic Reverse Saturation Impedance. Dynamic reverse impedance varies considerably, being, in general, directly proportional to the breakdown voltage of the particular diode. Table 2 shows typical values of reverse resistance (in the saturation region) for a few of the available types.

The saturation reverse resistance also varies markedly with bias current for a particular diode, being inversely proportional to current over a relatively large range near the "knee," but approaching a more constant value at higher currents.

Table 2. Silicon-diode Reverse Resistance

Diode	Reverse saturation voltage, approx.	Saturation dynamic resistance,* approx. ohms
1N401	1.8	5
1N405	4	10
1N202	12	50
1N207	34	280
1N210	58	750

* At nominal bias current.

21. Reverse Capacitance. Capacitance is not generally important in regulation applications, but reverse capacitance varies approximately as $1/\sqrt{V}$. Manufacturers usually give a nominal value at a particular voltage, commonly 6 volts. $C_{v_2} \approx C_{v_1} \sqrt{V_1/V_2}$, where C_{v_2} is the capacitance to be determined at a voltage V_2, and C_{v_1} is the nominal capacitance given by the manufacturer for a voltage V_1. Forward capacitance is rarely specified and may range from tens of micromicrofarads to thousands of microfarads depending on physical size of the diode and the forward current level; it is directly proportional to both.

22. Reverse-biased Switches. Diodes having abrupt voltage breakdown may be used as switches while biased only in the reverse direction. Operation in a high-impedance or "off" state is conventional, but the "on" or low-impedance state is

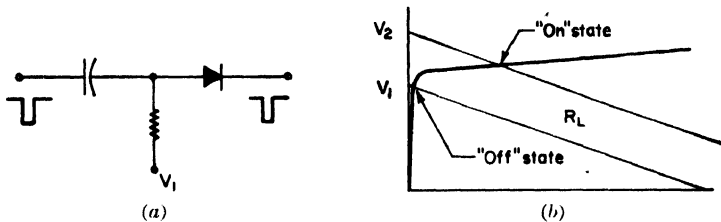



FIG. 21. Biasing of a reference diode for use as switch.

obtained by increasing the reverse bias so that breakdown occurs and a high current flows.

If low-voltage regions are avoided, a relatively low capacitance is obtained and, further, very low switching times can be obtained since very rapid changes (tens of millimicroseconds or less) in current may occur in the breakdown region of the characteristic. Figure 21 illustrates this application.

23. Temperature Coefficient of Saturation Voltage. This is usually positive and directly proportional to voltage, at voltages above approximately 4 volts. At lower voltages a negative temperature coefficient is observed. The exact zero-coefficient voltage varies somewhat with manufacturing-process details and is not constant except as a statistical mean for a particular diode type. However, the possibility of series-connecting two diodes having temperature coefficients which more or less cancel one another over a substantial temperature range does exist.

Dissipation Ratings. Currently available types can dissipate up to about ½ watt, but laboratory devices capable of tens of watts dissipation have been constructed and will be available in the future.

A-c Voltage Regulation. This can also be accomplished by using a double peak-

clipping circuit. Either single or multistage clipping can be employed, depending upon the desired degree of regulation, as shown in Fig. 22.

24. Silicon Point-contact Diodes. Silicon point-contact diodes are constructed as two important generic types for use as high-frequency microwave detectors and in the more conventional design as a fast-recovery type which is especially useful in high-speed computer circuits. In appropriate point-contact designs a low-capacitance diode having low storage time and a high rate of recombination of hole-electron pairs may be constructed which may be useful as a detector to 500 Mc or higher without special assembly design.

For use as a microwave detector or frequency converter the diode is normally constructed in a manner suitable for use as part of a transmission line or cable. These diodes vary considerably in design depending upon particular applications and the

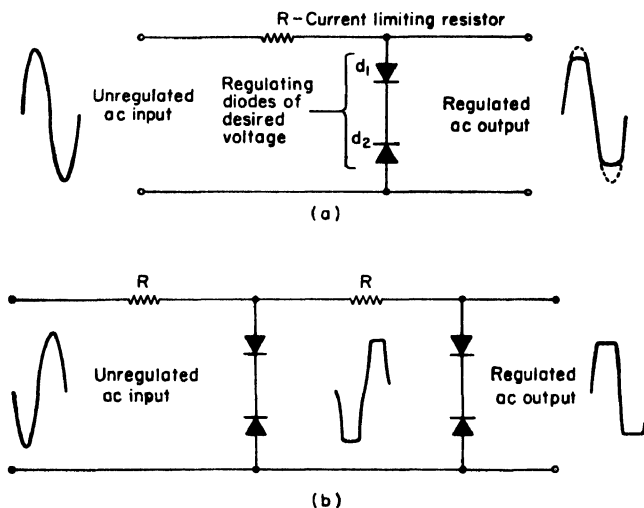



Fig. 22. A-c voltage regulation with breakdown diodes. (*a*) Single stage. (*b*) Two-stage arrangement.

characteristic impedance of the transmission line or mixer with which they are to be used. Certain specially designed types may have by-pass capacitance and lead-inductance features incorporated in the diode enclosure itself. See Fig. 18 for examples.

For use as a high-speed switching or computer diode the same characteristics make the silicon point-contact diode extremely useful where rapid reverse recovery is required. Recovery times on the order of hundredths of a microsecond may be realized under practical computer switching conditions. Extreme care is required in measuring transient responses of such short durations. These diodes may be designed to have low reverse leakage at 100 to 200 volts and several milliamperes, or more depending on particular design compromises, of forward conduction at 1 volt. They are normally assembled in small or miniature glass envelopes.

25. Germanium-junction Diodes. Germanium-junction diodes are made in alloy, grown, and diffused versions, the former being the most common, especially in the high-current types. They are made in a wide range of sizes, but most notable are the larger units intended for high current rectification and switching. Their most important characteristics are the high forward conduction, up to hundreds of amperes at ½ to 1 volt, and relatively low reverse leakage currents. The forward "knee" voltage occurs at about 0.3 volt in contrast to that of about 0.7 volt for silicon.

They are, thus, especially useful in low-voltage and/or high-current applications

where the low voltage drop and low dissipation are desirable. In smaller types, low- to medium-speed switching is possible, and high rectification efficiency may be obtained at all power levels with suitable types. In current designs, temperature limitations are, nominally, in the vicinity of 100°C, though there are basic reasons for expecting that this can be increased in the future.

Reverse voltage breakdown is less abrupt and more variable than with silicon; so germanium diodes are seldom used for reference or regulating purposes though they are useful in all other applications within their frequency range.

26. Germanium Point-contact Diodes. *Tungsten-whisker Types.* These diodes are representative of a family of germanium point-contact diodes using unplated whiskers, usually tungsten. These diodes have relatively low forward conductance, 1 to 25 ma approximately at +1 volt, and relatively high reverse leakage. The familiar 1N34 or 1N34A is more or less representative of this generic diode type.

Reverse capacitance of these diodes is relatively low. In certain design variations reverse recovery time is as low as 0.1 to 0.25 μsec, and such diodes are excellent for general-purpose use as high-frequency (to about 300 Mc) detectors and computer diodes. Depending upon the particular diode characteristics specified (forward and reverse conductance, recovery time, etc.) a wide variety of types are available.

27. High-forward-conductance Types. These diodes are known by a variety of designations, such as high conduction, low impedance, plated whisker, and bonded.

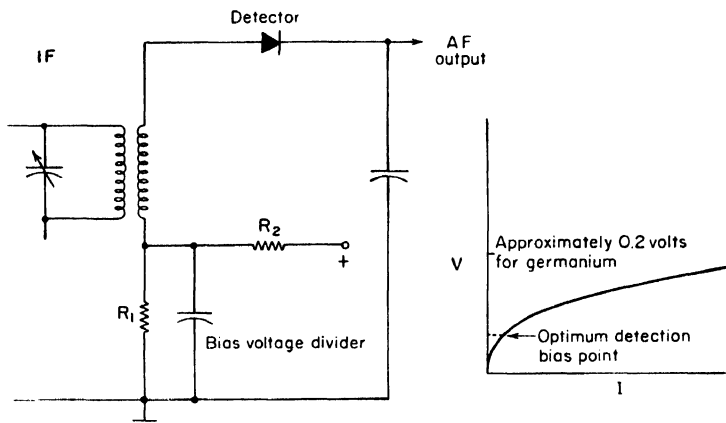



FIG. 23. Optimum low-level detector biasing.

They are made by plating the whisker with a metal such as indium or gold and subsequently heating the contact with the germanium die by electrical pulsing to melt the whisker plating and, locally, the germanium die. Recrystallization produces a pseudo-junction type of diode having greater area than before pulsing. This increased area contributes to higher forward conduction, which results from the formation of this crude junction, which also has a lower reverse leakage than a true point contact.

Diodes constructed in this manner have characteristics intermediate to point-contact and alloyed-junction diodes. Forward conduction is usually in the range of 25 to 500 ma at 1 volt, and reverse leakage current may approach that of junction diodes.

Reverse capacitance and recovery time are generally somewhat greater than those of the true point contact. Storage time is also higher, though this is an advantage in diode amplifiers or other applications where this characteristic may be exploited.

Applications. These diodes are generally applicable where point-contact types may be, though not quite so useful in high-frequency and high-speed computer applications. The more abrupt transition into (higher) forward conduction makes them very useful as clamps and low-power rectifiers, but some forward bias is desirable if full advantage of this characteristic is to be taken in detector service (see Fig. 23).

SEMICONDUCTOR DIODE REVERSE CHARACTERISTICS

28. Reverse Current. Reverse resistance in semiconductor diodes is a nonlinear function of voltage and obeys different laws in various voltage regions. These regions are shown graphically for a germanium-junction diode in Fig. 24. In region

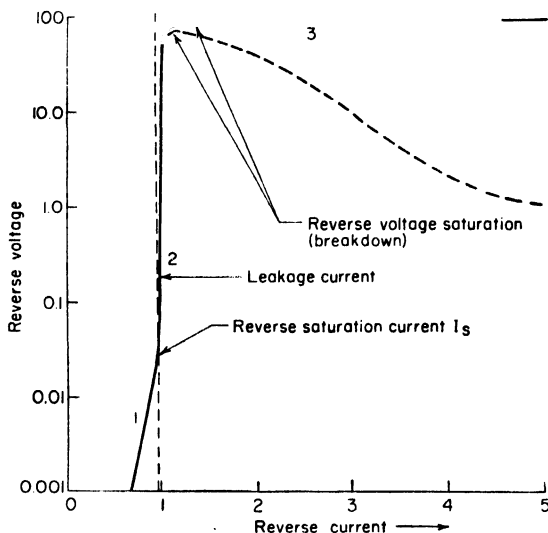



FIG. 24. Reverse current in a semiconductor diode.

1 of Fig. 24 up to a few tenths of volts (approximately the barrier potential for the material and temperature involved) the current is composed of thermally generated minority carriers which have diffused to the junction region and are collected there. The charge carriers are hole-electron pairs generated when electrons gain enough thermal energy to escape from their parent atoms, leaving a net positive charge called a hole which may be filled by another electron. For a given temperature there are only a limited number of thermal carriers available and a reverse current saturation results at about 0.3 and 0.7 volt, respectively, in the cases of germanium and silicon, respectively.

FIG. 25. Reverse resistance of a semiconductor diode.

At higher voltages, reverse current still increases (region 2 of Fig. 24) but does so more slowly. This is due to surface or body leakages, which are usually nonohmic, though they may be ohmic over a fairly wide voltage range in some cases. At still greater reverse voltages, the reverse current increases very rapidly (region 3 of Fig. 24) because of one or more diode-breakdown phenomena. In certain diodes, this breakdown is quite abrupt and a negative resistance region may be observed—perhaps leading to excessive dissipation and possible catastrophic failure of the diode.

In the region of normal reverse operation, the reverse junction current may be approximately given by $I = I_s + V/R$ (Fig. 25). In the cases of some junction diodes and most point-contact types, the incremental resistance is nonohmic and

linear extrapolations cannot be used in predicting reverse current at one voltage from that at another. For a perfect junction, this resistance would be infinite.

I_s is a function of temperature and can be predicted fairly accurately if known for one temperature. The incremental leakage resistance is commonly a function of temperature but varies considerably depending upon the individual device and type of device.

For germanium: $$I_{sT} \approx I_{sT_1} \times 2^{\frac{T_2 - T_1}{10}}$$

For silicon: $$I_{sT} \approx I_{sT_1} \times 2^{\frac{T_2 - T_1}{5}}$$

where $I_{sT_2} = I_s$ at temperature T_2.

The behavior of reverse current in semiconductor diodes is similar to that of collector or emitter reverse currents in transistors.

29. Reverse Breakdown (Voltage Saturation). Reverse voltage breakdown may be caused by one or more factors. In general, if diode temperature is limited to rated values, no impairment results from excursions into the breakdown region. The most important causes of reverse voltage saturation or breakdown follow.

Zener or Field Emission. In Zener breakdown electrons are torn from their atomic bonds by the high electric field across the junction. These electrons may then cause secondary emission, and a cascade or avalanche type of breakdown may then be involved as well.

Avalanche. Avalanche breakdown occurs when available electrons (as thermal charge carriers) gain enough energy in traversing the junction to cause secondary emission and a cascade or avalanche results. This mechanism is similar to that of ionic glow discharges in gases. In nonuniform junctions (as in point-contact devices) avalanches may occur first in localized high-field regions, progressing to others as the field increases and a "soft" or rounded breakdown characteristic results.

Punch-through. As reverse voltage is increased, the junction-depletion layer widens and, if no other breakdown processes intervene, may eventually, at a lower voltage, include the entire base or die thickness. Further increase in voltage cannot then cause a further increase in junction thickness and the electric field increases more rapidly until it induces breakdown.

Surface Breakdown. As very high electric fields exist across the surface, as well as in the interior of the diode, breakdown or leakage may occur at the boundary of the junction and its ambient. Conductive contaminants if not completely eliminated may contribute to high reverse leakage currents or a lower breakdown voltage. High-voltage low-leakage diodes are often coated with or encased in materials having good dielectric qualities to minimize these effects.

It is frequently very difficult to say with certainty which of these mechanisms is causing reverse breakdown or saturation in a particular diode or transistor, as failure may be caused by any one or a combination of several of these mechanisms if critical dissipation levels are exceeded. In certain cases localized burnout and progressive deterioration may result if nonuniform junction temperature occurs.

Thermal Breakdown or Runaway. The reverse current in semiconductor devices varies in an exponential manner with temperature. Further, the temperature depends upon the dissipation, which is a direct function of reverse current and voltage.

As voltage is increased (and with it reverse current) the dissipation rises and with it junction temperature, and at some critical dissipation and temperature this regenerative situation leads to an unstable or runaway condition. This critical point is that at which

$$\frac{dT}{dP}\frac{dP}{dT} > 1$$

where dT/dP = rate of increase in temperature with dissipation
dP/dT = instantaneous rate of increase of dissipation

For a constant reverse voltage $dP/dT = V(dI_R/dT)$. Then the temperature at which $dI_R/dT > (1/V)(dP/dT)$ is the runaway point. Thus, if the dP/dT and dI/dT characteristics of the individual device are known, its runaway temperature can be determined. These characteristics are frequently available from the manufacturer and are implicitly given in his ratings. The former is sometimes quite uniform for a type of device, while the latter varies widely with individual units. For nearly ideal junctions at voltages below breakdown, I_R may be used to compute approximate critical or runaway temperatures. Typical forms of these characteristics are shown in Figs. 26 and 27. Note that widely different curves (only two of which are shown) for dI_c/dT may result, depending upon the temperature relationships of the various components of the reverse current.

FIG. 26. Junction temperature rise vs. dissipation; air cooling.

FIG. 27. Typical form of junction reverse current vs. temperature.

DIODE RECOVERY TIMES

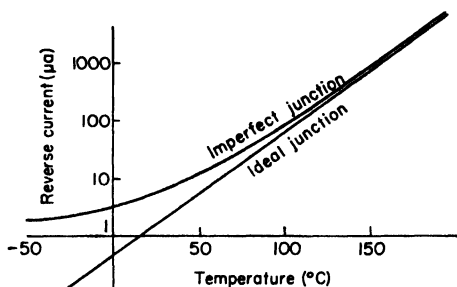
When a semiconductor diode is switched from the "on" to "off" state, the transition from a high (forward) current to a low reverse current is not instantaneous but requires a finite period of time. Similarly, a transition from "off" to "on" is not instantaneous, though this time is often short enough that it can be considered negligible.

As a relatively long time is required for such switching transients to disappear completely, it is conventional to term the period between 10 and 90 per cent of final current or voltage values as the "switching" or "recovery" time, or, in appropriate cases, "turn-on" or "turn-off" time. Recovery to a specified reverse-current or resistance (large-signal) value is also a commonly measured time.

It may be desirable to test the diode in an actual circuit and determine the time required for recovery to a specified current or voltage, implicitly to a minimum large-signal resistance. The test circuit is usually the actual circuit or is related to that in which the diode is ultimately to be used. It is also possible that a rectification-efficiency test at a suitable frequency will correlate with actual switching-time measurements or, perhaps, be acceptable when time cannot be measured simply or accurately.

30. Forward-recovery or Turn-on Time. This is the time required for an "off"-biased diode to switch to an "on" or high-conduction state. The diode under test is usually biased off by a fixed voltage and is pulsed on by a high-impedance or current source. The voltage across the diode is observed on a suitable oscilloscope, preferably one with a time-calibrated sweep. The oscilloscope's vertical amplifier must have a transient response considerably shorter than the diode recovery time to be measured.

The voltage waveform across a typical computer diode is shown in Fig. 28. A commonly used test circuit is also shown. The turn-on time is commonly defined as the time required for the voltage to return to less than a specified value. Alternatively, time is not actually measured, but a maximum peak forward voltage may be specified. In the case of certain high-speed diodes, there may be no visible overshoot, indicating that turn-on time is less than the oscilloscope's response time.

31. Reverse-recovery or Turn-off Time. This is the time required for an "on"-biased diode to switch to an "off" or high-impedance state. The diode under test is

normally biased on by a fixed current and is pulsed off by a low-impedance pulse generator. The time required for reverse current to recover to an arbitrary fraction of the final value or, commonly, to a fixed value is measured on a calibrated oscilloscope. The wave form of reverse current, as measured across a series resistance, and a suitable test circuit are shown in Fig. 29.

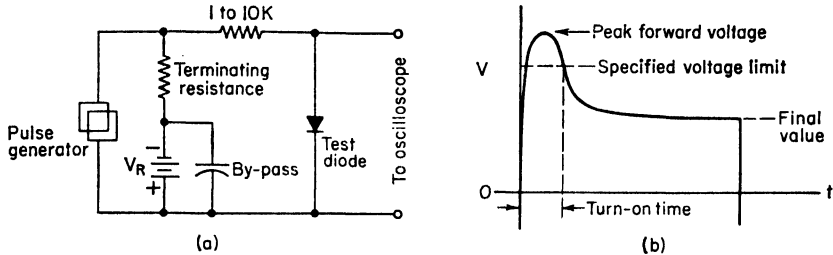



FIG. 28. (*a*) Forward-recovery-time test circuit. (*b*) Diode forward current during pulse conduction.

The pulse of reverse current is a result of storage of minority carriers in the base region, which must be collected at the oppositely biased region. In cases where forward dissipation is significant, thermal recovery is also involved. When times of less than 0.1 μsec are to be measured, great care must be given to designing the test circuit so that diode and not circuit response is measured. Major difficulties may be associated with:

1. The signal source. A fast rise time is required of the applied step function. Suitable pulse or square-wave generators or relay contacts may be used.
2. The scale-switching diode must have low capacitance and a short recovery time, preferably much less than that of the diode under test.
3. The oscilloscope amplifier and display circuits should have a sufficiently rapid response.

FIG. 29. Diode reverse recovery current.

All these factors may introduce time constants on the order of 10 mμsec or more so that recovery times of 25 to 100 μμsec can be measured only with some difficulty and potentially significant errors. When recovery times of less than about 50 μμsec are to be measured, a rectification-efficiency test should be considered as the circuit is simple and dependent mainly on the test diode's response, which usually correlates well with performance in the switching circuit.

32. Diode Detection or Rectification Efficiency. A test of diode rectification efficiency is very useful in evaluating the utility of a diode as either a detector or rectifier if made under conditions representative of its intended use. It may also be invaluable in estimating performance of a fast-recovery switching diode. Test conditions of importance are the use of (1) signal source of a suitable frequency, (2) output level,

(3) a filter output circuit realistic in terms of the actual operating circuit, and (4) suitable indicating meters.

A basic rectification-efficiency test circuit is shown in Fig. 30. The filter capacitor and load resistance are chosen for the desired time constant and power consumption. The meter may be a d-c VTVM or a simple resistance-current meter combination, sometimes used as a load as well. In certain detector applications a modulated input

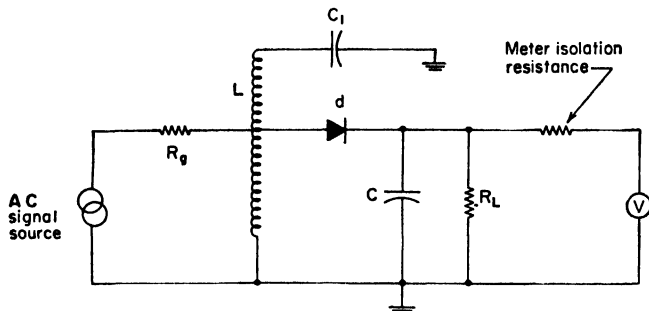



FIG. 30. Basic circuit for testing rectification efficiency.

signal and an a-c voltmeter are used. A high-Q tank circuit may be used to filter the input current so that it may be more readily measured, when the wave form of diode current is not readily determined or will load the source unduly.

In evaluating computer diodes, a frequency of tens to hundreds of megacycles is often used, and where desired the output capacitance may usually be made as low as that of the circuit in which the diode is to be used. This is a relative advantage compared with the classical recovery-time-measuring circuits discussed where many large time constants limit measurements. Correlation of rectification efficiency with time measurement can usually be made in the 50- to 150-mμsec range and extrapolated to considerably lower values.

TEMPERATURE EFFECTS

33. Cooling of Power Diodes. Certain diodes do not have sufficient cooling capacity inherent in their package design, and these must be attached to cooling structures

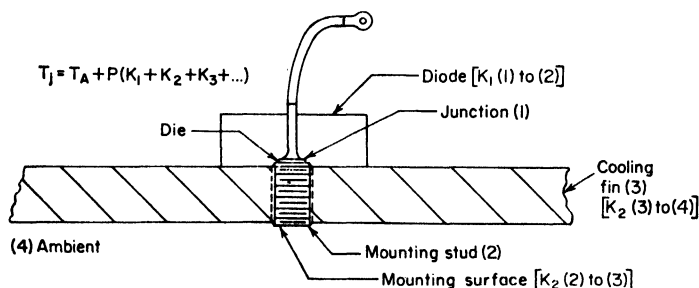



FIG. 31. Temperature gradients involved in power diode cooling.

or have a current of air forced past them. It is important to remember that internal-temperature-gradient coefficients of dissipation are limiting factors over which only the manufacturer has control and that the user must determine actual dissipation in his application and provide sufficient cooling capacity so that maximum rated junction temperature is not exceeded.

Junction temperature is given by the formula in Fig. 31. Temperature-gradient coefficients must be obtained from the manufacturer and determined by the user for

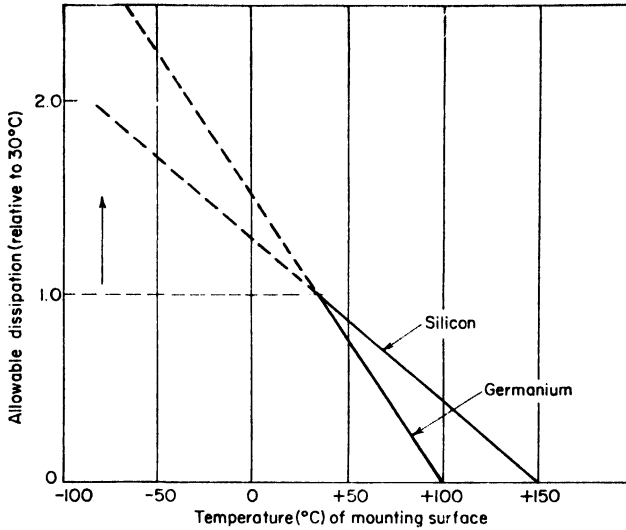



FIG. 32. Temperature derating of diodes.

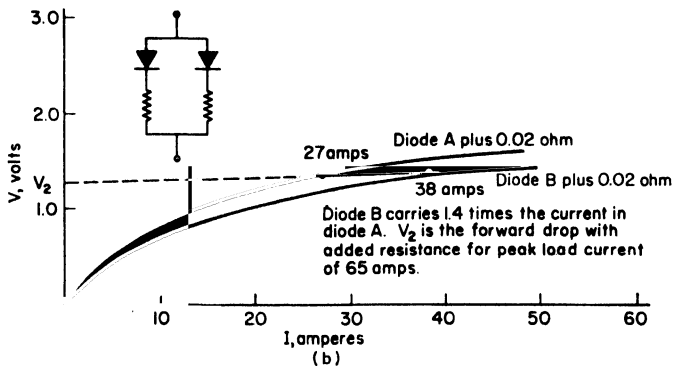
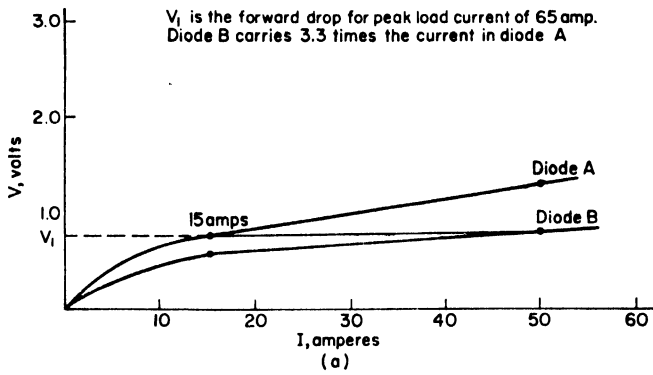



FIG. 33. Improved uniformity of diode forward characteristics when series resistance is added.

the cooling structure to be used. This rise in temperature is analogous to a rise in voltage when a current is fed to a series of resistances. In this case, dissipation and temperature-gradient factors are analogous to current and resistance, respectively. The various coefficients are physically associated with diode construction and mounting in Fig. 31.

The temperature-gradient coefficients may be nonlinear and vary with temperature. This is especially so in the case of fluid convection and radiation cooling (see Fig. 26), and static values at arbitrary or limiting values of junction and ambient temperatures are often quoted in manufacturers' literature.

Figure 32 illustrates the typical form of power-diode temperature derating. Actual maximum temperature limits are quite variable depending on type, as are the slopes of the derating curves. Some manufacturers permit linear extrapolation to low temperatures, but in some cases constant maximums may be established.

SHUNT AND SERIES OPERATION

Occasionally, because of the limitations of ratings, it is necessary to operate diodes in shunt to obtain higher current capacity or in series to obtain higher inverse voltage ratings. In such cases, the diode forward or reverse conductances must be carefully matched and remain matched during their life, or unequal sharing of dissipation results, leading to the failure of some or all of the diodes involved.

34. Shunt Operation. Where two or more diodes are shunt-operated to increase their allowable output current, voltage across them is equal, and if current flowing is

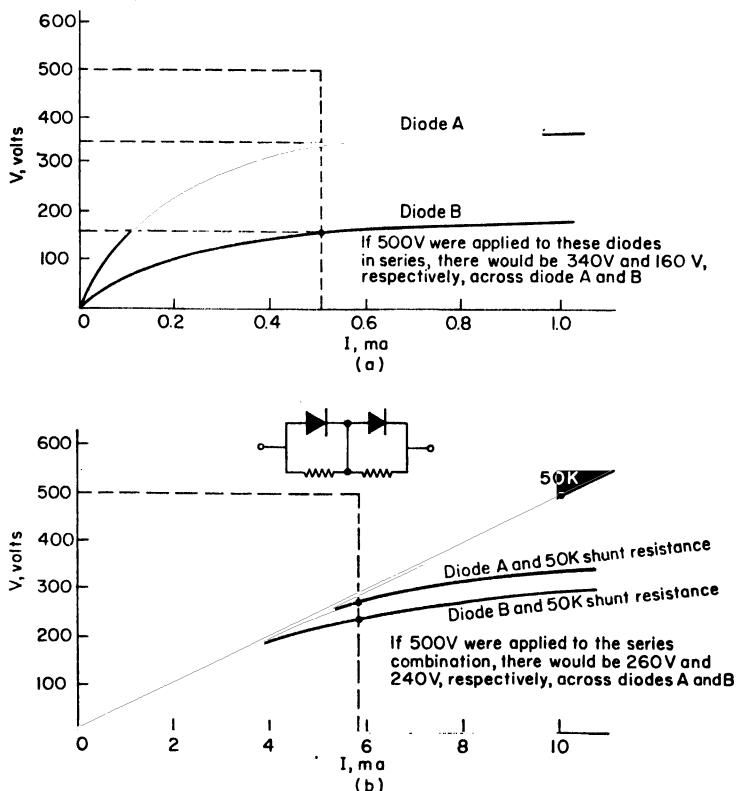



FIG. 34. Improved uniformity of operation with series diodes plus shunt admittance.

not also equal, the diode having the higher forward conductance carries a higher fraction of the output current and must dissipate more heat than the other. In many diodes, the higher temperature of the better diode further increases its conductance and its temperature, leading to early failure if it cannot carry almost all the load alone.

When the added voltage drop may be tolerated, diodes having poorly matched or nonuniform forward characteristics may be shunt-operated if a resistance is inserted in series with each diode. If this resistance is five to ten times greater than the internal (incremental) diode resistance, the effective uniformity of the diode characteristics is improved and parallel operation with uniform sharing of load current is possible.

Certain silicon diode types may be exceptions, as their forward resistance increases with temperature, tending to compensate for nonuniformities, and little if any series resistance may be needed. Where dissipation in the series resistances is excessive, inductive ballasts may be used instead where a-c or pulsed voltages are applied.

Figure 33 shows the forward characteristics of two germanium diodes having different incremental resistance above the knee with and without added series resistance.

35. Series Operation. When two or more diodes are operated in series to obtain higher breakdown voltage, the current in all diodes is equal, and unequal distribution of reverse voltage results if the reverse characteristics of the diodes are not uniform throughout life.

Where the increased current can be tolerated, resistances may be added in shunt with each diode. Where dissipation in the shunt resistances may be excessive, capacitive voltage dividers may be used instead when a-c or pulse voltages are applied.

The resistance values should be chosen so that the resistor current at peak voltage is much greater than that in the diode, preferably five to ten times.

Figure 34 shows the characteristics of two diodes with and without their shunt resistors. Note the improved division of the applied voltage.

FIG. 35. Photocurrent characteristics of a typical selenium photovoltaic diode.

PHOTODIODES

Photodiodes are two-element, light-sensitive devices which (1) generate a voltage or (2) undergo a change in effective resistance when exposed to light. Accordingly they may be termed photovoltaic or photoresistive, respectively. Many semiconductor photodiodes may be operated in either manner, *e.g.*, selenium is useful in a light meter as a self-generating photocell but may also be used with external bias as a photoresistive cell, while cadmium sulfide is used only as a sensitive photoresistor.

36. Photovoltaic Diodes. Photocurrent characteristics of a typical selenium photodiode are shown in Fig. 35. Note that this is a relatively low internal-impedance device and that output voltage tends to saturate at high light levels and is on the order of 0.25 volt. Silicon and germanium photovoltaic diodes will deliver approximately the same output voltages under load conditions, though open-circuit voltages are in the 0.25- to 0.5-volt range. Multiple stacks may be used to obtain higher voltages and currents for various applications. Spectral response of a selenium photocell is shown in Fig. 36. The peak response corresponds to that of the human eye.

37. Photoresistive Diodes. Photo characteristics of a typical germanium photoresistive diode are shown in Fig. 37 (in the case for a small-area germanium-type 1N77B). Voltage and current output swings can be readily determined from these curves if bias-voltage, load-line, and light-input excursions are known.

Dark current is equivalent to junction-leakage current, as discussed in Sec. 28, and at normal temperatures is lowest for silicon and progressively higher in germanium or selenium and other compounds.

Cadmium sulfide is notable for its high sensitivity, which is due to an inherent current-multiplication effect. Spectral sensitivity of germanium and silicon diodes is given in Fig. 36.

FIG. 36. Spectral response of semiconductor photodiodes.

38. Frequency Response. Frequency response of photocells depends upon their construction, point-contact types having good high-frequency response, while large-area junction types are relatively poor. Certain selenium cells have very slow transient recovery, *e.g.*, those having time constants as high as 1.0 to 100 sec or more.

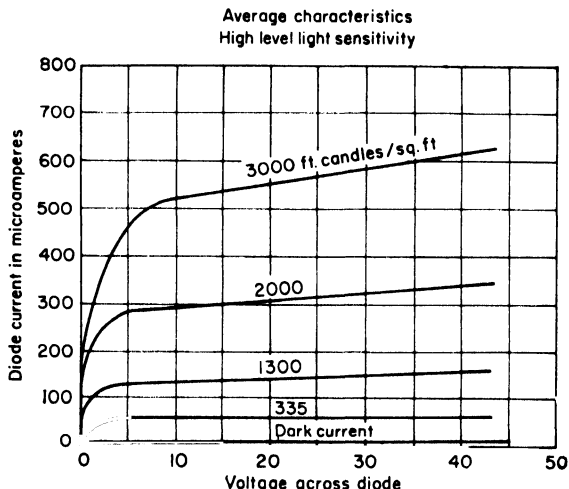



FIG. 37. Photo characteristic of a typical germanium photoresistive diode.

39. Temperature Stability of Photodiodes. Temperature stability of photovoltaic cells is on the order of about 0.1 to 5 mv per °C. In the case of photoresistive types silicon and germanium are relatively stable though dark current varies in the same manner as reverse leakage (see Sec. 28). Other types have been somewhat less stable and may be unusable with d-c amplifiers when exposed to large temperature variations if the dark current is a significant fraction of the light current.

CHAPTER 10

TRANSISTORS

By Seymour Schwartz[1]

In many ways the development of the transistor has paralleled that of the vacuum tube as the latter is known today.

In 1883, Edison discovered that a unidirectional current would flow through a two-electrode evacuated enclosure. Not until 1904 was any practical use made of this discovery. Then Fleming applied it to the detection of wireless signals. In 1906, De Forest added a third or control electrode to the simple Fleming valve, thereby making a simple rectifier into an amplifier and, as a consequence, into an oscillator.

The discovery of the fact that metals in contact with certain "semiconductors" would conduct current asymmetrically seems to go back to Munck in 1835. This rectifying action was utilized by Pickard, who, in 1906, patented the crystal detector. His "cat-whisker" device was widely used until it was replaced by De Forest's triode. Then the crystal diode virtually disappeared until it was resurrected and improved during the Second World War.

Just as De Forest's grid made possible totally new devices and applications, so the addition of a third electrode to a cat-whisker crystal detector has brought about a vast and important development.

First announced in 1948 by Brattain, Bardeen, and Shockley, the crystal detector with two cat whiskers became the transistor. Theoretical work by Shockley and many others plus experimental work of countless engineers and scientists have brought the device to the point where it is exceedingly useful. The art is already complex but still fluid so that the transistor picture today may not be at all what it will be a few years hence.

So rapid has the development been that the first type of transistor—the point-contact type—has already been put aside in favor of newer and better types.

1. Basic Transistor Theory. Certain elements have electrical resistivities intermediate between good conductors and good insulators. In general the resistivity of these semiconductors increases with increasing temperature.

A transistor is a device made from such a semiconductor, notably silicon or germanium, in which certain impurities have been incorporated. Some impurities, such as arsenic or antimony, known as *donors* can act as suppliers of electrons; others such as gallium and indium and known as *acceptors* can act as suppliers of positive charges.

Materials having an excess of negative charges (electrons) are called *n-type* materials; and those which have an excess of positive carriers are called *p-type* materials.

In *n*-type materials, the *majority* carriers are electrons; in *p*-type materials the majority carriers are *holes*. Conversely in *n*-type materials the *minority* carriers are holes and, of course, in *p*-type materials the electrons are minority carriers. In transistors it is the minority carriers that are controlled and provide the useful amplification.

Current through the device is considered as being conducted by the electrons and by holes which are created when an electron breaks free from the crystal in which it normally resides. Conduction in transistors cannot be explained simply by the

[1] Transistor Applications Inc. The author is indebted to A. W. Carlson of Transistor Applications Inc. and Robert McMahon of Lincoln Laboratories, MIT, for their assistance and contributions.

Table 1. Transistors, Types and Characteristics*

Device property \ Process	Doped grown	Rate grown	Surface barrier	Fusion
Region resistivity:				
Emitter	Very low	Moderately low	....	Very low
Base	Moderately low	Moderately high	Moderately low	Moderately high
Collector	High	Moderately high	....	Very low
Region size:				
Emitter	Large	Large	....	Thin
Base	Thin	Thin	Very thin	Thin
Collector	Large	May be quite thin	....	Thin
Limitations on alpha	Carrier transport and surface recombination	Emitter efficiency, carrier transport, and surface recombination	Emitter efficiency	Carrier transport, side diffusion, and surface recombination
Limitations on low r_b'	Spreading resistance from point (poor ohmic)	Spreading resistance from point and moderately high base resistivity (poor ohmic)	Very thin web	Thin web
Limitations on max collector voltage	Bulk and/or surface breakdown	Bulk and/or surface breakdown	Surface breakdown and/or punch-through	Punch-through
Phenomena limiting R_{re} (dynamic)	Surface only	Surface and some space-charge widening	Surface and space-charge widening	Space-charge widening and some surface
r_c	Very high	High	Low	Moderately high
Alpha crowding	Good	Probably poor	Not of concern due to power limitations	Good
Collector series resistance R_{cs}	High	Moderate to low	Very low	Very low
Power capabilities	Only fair	Only fair due to alpha crowding	Very low due to small areas	Good for geometry reasons
f_{co}	High 5–10 Mc	High 5–10 Mc	Very high (Ge) 50–150 Mc	Moderate 1.5–5 Mc
Collector capacitance	Low due to high collector resistivity and grated junction	Low due to high collector resistivity	Low due to small size	Moderate high due to low collector resistivity, larger collector than emitter and step jn.
Emitter inverse voltage	Low due to bulk breakdown	Moderate low due to bulk breakdown	Low due to surface breakdown	High, same as collector
Max collector voltage	Ge, 30–70 volts; Si, 80–120 volts	Low	ca. 6 volts	Ge, 30–70 volts; Si, 30–200 volts
Emitter-to-collector shorting capacitance	Appreciable due to base contact overlap	Appreciable due to base contact overlap	Very low	Very low
Emitter floating potential	High (due to base channel)	High (due to base channel)	No such effect	No such effect
Collector low-voltage saturation characteristic	Poor due to high R_{cs}	Probably moderate to good	Good	Good

* Courtesy of Semiconductors Division, Hughes Aircraft Co., Los Angeles.

movement of electrons, and the hole concept of the conduction theory seems well established.

2. Point-contact Transistor. The first type of transistor, now outmoded, was the point-contact type consisting of a slab of the semiconductor called the *base*, and two cat whiskers, called the *emitter* and the *collector*. In a general way these electrodes correspond to the grid, the cathode, and the plate of a triode vacuum tube, but the

correspondence is not complete or exact. The transistor output characteristics resemble those of a pentode vacuum tube.

3. Junction Transistor. Nearly all the transistors in use today are of the junction type in which, for one example of construction, a sandwich is made of a *p*-type material between two slices of *n*-type material or the converse, that is, a slice of *n*-type material between *p*-type slices. In either case the middle material forms the base, and if this is of *n*-type material, then the device is called a *p-n-p* transistor and if the base is *p*-type, then the device is an *n-p-n* transistor.

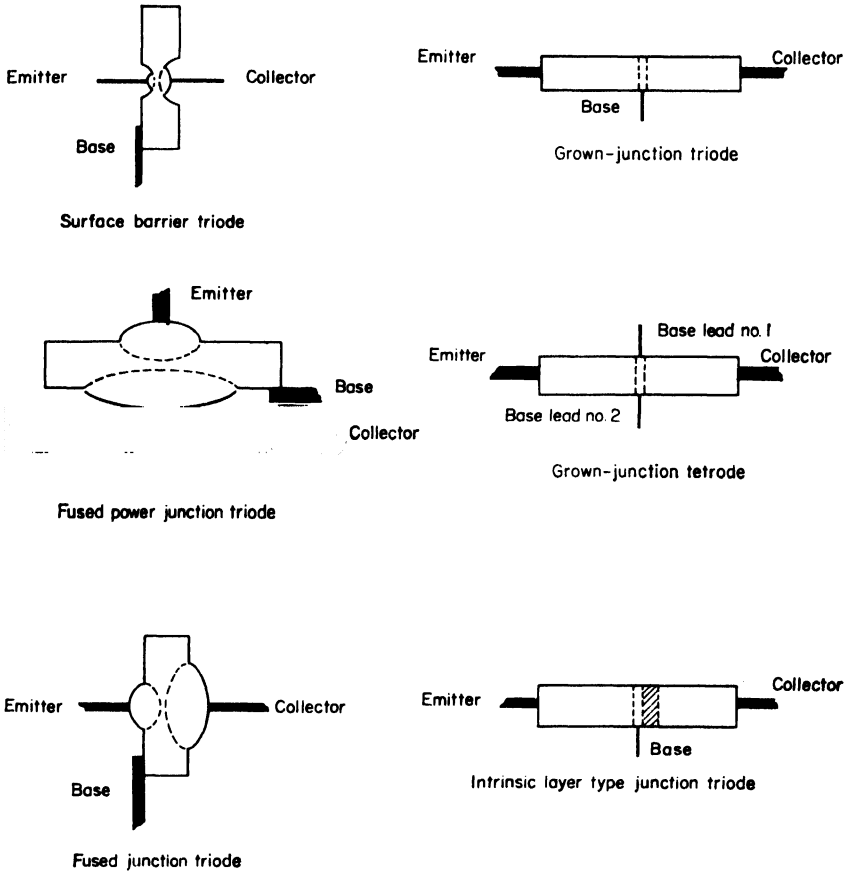



FIG. 1. Transistor geometry.

4. Transistor Characteristics. Because the transistor is small, because no stand-by heat is required to make it function when needed (there is no heated cathode or filament), because of the elimination of microphonics, because there is considerable power gain, and because there is very little reason for it to wear out, the transistor has become an electronic tool with very great possibilities.

There are some disadvantages, however. One is the fact that, even though biased in the reversed direction, a small steady current flows from an electrode to base whether there is any input driving power or not. The disadvantage comes not from the loss of efficiency or wastage of power or production of heat to be dissipated but

Table 2. General Distinguishing Characteristics of Junction Transistors*

Transistor groups	Major characteristics that distinguish type numbers	Most commonly used types	Typical range of max collector dissipation, mw	Typical range of max collector voltage breakdown, volts	Typical range of max I_{co}, μa	Typical range of C_c, μμf	Typical range of β (α_{cb})	Typical range of $f_{\alpha_{co}}$, Mc
Low-frequency low-power	Low α, β; medium α, β; high α, β; low noise	*p-n-p* fused Ge, *n-p-n* grown Ge	50–150	30–50	3–10	3–20	30–100	0.5–1.5
Low-frequency low-power high-temperature	Low α, β; medium α, β; high α, β	*n-p-n* grown silicon, *p-n-p* fused silicon	50–150	30–50	<0.1	3–10	15–50	0.3–2
Low-frequency high-power	Medium p_c, high p_c, value of β at high I_e	*p-n-p* fused Ge	100–20,000	20–60	20–2,000	25–400	5–50	0.01–0.2
High-frequency low-power	Medium $f_{\alpha_{co}}$; high $f_{\alpha_{co}}$; value of $R_{b'}C_c$	*p-n-p* fused Ge, *p-n-p* surface barrier Ge, *n-p-n* tetrode Ge, *n-p-n* grown Ge	20–50	6–30	1–10	2–15	30–100	1–100
High-frequency low-power high-temperature	Medium $f_{\alpha_{co}}$; high $f_{\alpha_{co}}$; value of $R_{b'}C_c$	*n-p-n* grown silicon, *n-p-n* grown tetrode silicon	20–50	6–30	<0.1	2–10	15–50	1–50

For meaning of symbols, see Sec. 9.
* As used in sine-wave circuit applications.

from the fact that this continuous current flow is strongly temperature-dependent, approximately doubling for each 10°C rise in temperature. If not taken into account in circuit design, these variations with temperature may cause trouble, in d-c amplifiers as one example. Another disadvantage is the dependence of collector capacitance on variations of collector voltage and base current.

Transistor fabrication is still an art, and some manufacturers continue to rely on selection techniques in sorting production into type-number categories. It must be realized, however, that even today some of the most important tube parameters are not held in manufacture to tolerances closer than about 20 per cent, and numerous parameters are not controlled at all.

5. Transistor Types. The literature and circuitry of transistors are already complex. Numerous types of the devices are available, their names and characteristics, advantages and disadvantages being functions of their structure. Only time will tell which of the large number now at hand will survive and which will fall by the wayside. Six of the common types are shown diagrammatically in Fig. 1, and Table 2 gives the distinguishing characteristics. In general the majority of these types of transistors may be classified into one of five major categories: (1) low-frequency, (2) high-frequency, (3) power, (4) low-frequency high-temperature,

and (5) high-frequency high-temperature types. This chart also points out the major characteristics that normally distinguish the units listed. It is of interest to note that, in general applications, most of the transistors in each of the respective categories are fairly interchangeable. Other information listed includes the types of transistors most commonly used and typical ranges of some of their more important characteristics.

6. Basic Operation. In a very elementary way the principle of transistor action may be explained as follows. Each of the end regions (Fig. 3), for example, forms with the base a diode rectifier with considerably greater resistance if biased in one direction than if biased in the opposite direction. The diodes are independent of each other except as the conduction of one is affected by phenomena taking place in the base. The resistivity between one end region and base is a function of the bias voltage between these elements and also a function of the voltages applied between base and the other end region.

Power amplification is secured because virtually the same current flows in the output as flows in the input, but the output impedance is higher than that of the input so that the I^2R available output power is greater than that in the input.

Basically the transistor is current-operated in contrast to a vacuum tube, which is voltage-operated. A current gain as high as 100 may be secured in some circuit and transistor configurations.[1]

7. Basic Circuits. Three basic circuits are possible, as shown in Fig. 2, depending upon which two electrodes are the input and which make up the output. One electrode or terminal is common to both input and output.

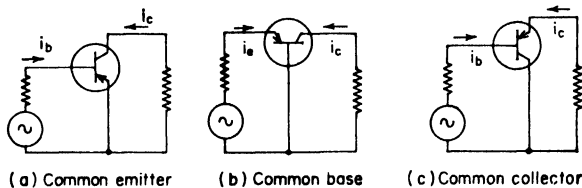



FIG. 2. Basic transistor circuits.

The *common-emitter* circuit is most widely used. It corresponds to the conventional grounded-cathode vacuum-tube circuit. The *common-base* circuit corresponds to the grounded-grid tube circuit; and the *common-collector* circuit resembles the cathode-follower type of tube circuit.

The common-base circuit has low input and high output impedances, current gain less than unity, and is a good r-f amplifier. Because of the low power gain it is not often employed.

The common-emitter circuit has high current gain, greater power gain than the grounded-base circuit, higher input and lower output impedances than the common-base circuit. The common-collector circuit has considerable current gain and an input impedance considerably higher than the output impedance. These relative characteristics may be summarized as follows:

	Current gain	Power gain	Input impedance	Output impedance
Common base	Low*	Intermediate	Lowest	Highest
Common emitter	High	Highest	Intermediate	Intermediate
Common collector	High	Lowest	Highest	Lowest

* Normally less than unity.

[1] For a more detailed and accurate explanation of the basic phenomena the reader should consult the references at the end of this chapter or D. G. Fink, Transistor Electronics, "Standard Handbook for Electrical Engineers," 9th ed., McGraw-Hill, 1957.

8. Junction Transistor Action. In Fig. 3 consider a junction transistor as being made of three slabs of material, a *p*-type base in the center with *n*-type emitter and collector materials on either side of the sandwich (grown-junction type).

Consider now the situation at the interface of the *p-n* junction (on the right). Without any bias potential being applied, the *majority* carriers (electrons in the *n*-type and holes in the *p*-type material) traverse the junction (*diffuse*) in equal numbers per unit of time. Thus there is no net current flow.

Now, if the *p*-type material, as shown in Fig. 3, is biased positive with respect to the base, carriers, positive and negative, move across the junction and a current made up of the sum of the carriers flows. The hole current, however, is greater than the electron current because the resistivity of the *p*-type material is lower. In the *n*-type material the holes injected across the interface are *minority* carriers. Because of the relatively enormous numbers of majority carriers in the *n*-type material, the holes are not subjected to any very great individual potential and they slowly diffuse through the material, recombining with negative carriers, with the result that the hole current gradually decreases. If the *n*-type material is thin so that most of the holes can get across it before recombining, the decrease in hole current is not so great.

FIG. 3. Geometry and bias conditions for grown-junction transistor.

If the junction is biased in the other direction, so that the *n*-type material is positive, electrons move away from the junction. The *p*-type material will now be negative and the holes move away from the junction on its side of the interface. With such a bias, very little current flows in either direction, perhaps only 1/100,000 of that flowing in the forward bias direction. Some current, however, always flows because of thermally energized carriers. This is the *saturation current* I_{co}.

The action at the other junction, the *n-p*, is exactly the opposite of that just described. The difference between the two junctions lies in their relative resistivities. The minority carriers diffuse through the material of the higher resistivity and the forward current direction is that in which the higher-resistance material is attractive to minority carriers. Thus the *n*-type terminal will be negative and the *p*-type terminal positive when the junction is biased for forward current flow.

9. Definitions. The following definitions are those of the principal electrical quantities or parameters encountered in transistor literature and employed in the remainder of this discussion.[1]

α = small-signal common-base short-circuit current gain $= \left.\dfrac{i_c}{i_b}\right|_{V_c=0}$

β = small-signal common-emitter short-circuit current gain $= \left.\dfrac{i_c}{i_e}\right|_{V_c=0}$

I_c = collector current
I_e = emitter current
I_b = base current
I_{co} = collector current when the collector is reverse biased and the *emitter* circuit is open at a given value of collector voltage
I_{co}' = collector current when the collector is reverse biased and the *base* circuit is open at a given value of collector voltage ($\doteq \beta I_{co}$)
V_c = voltage, collector to common terminal
V_e = voltage, emitter to common terminal
V_b = voltage, base to common terminal
f_α = frequency at which alpha is 0.707 of its low-frequency value (alpha cutoff frequency)
f_β = frequency at which beta is 0.707 of its low-frequency value (beta cutoff frequency)
C_c = capacitance between collector and base
r_e = small-signal emitter resistance

[1] For a more complete list of quantities and parameters see IRE Standards on Semiconductor Symbols, *Proc. IRE*, **44**, 934, July, 1956.

r_b = small-signal base resistance
r_c = small-signal collector resistance
r_m = mutual resistance (product of alpha and collector resistance)
p_c = collector power dissipation
r_b' = high-frequency base resistance

10. Equivalent Diode Circuits. It is often convenient to picture the transistor as shown in the diode equivalent circuit of Fig. 4. The collector-base and emitter-base junctions are represented by diodes with the collector diode shunted by a current generator αI_e. A resistance r_b represents the resistance of the base region. The circuit is shown for a *p-n-p* transistor; for an *n-p-n* transistor the diodes, voltages, and currents would be reversed.

The transistor is normally biased as shown with the collector diode biased in the reverse direction where it shows a high impedance, and the emitter diode is biased in the forward direction where the impedance is low. Ideally the emitter current would

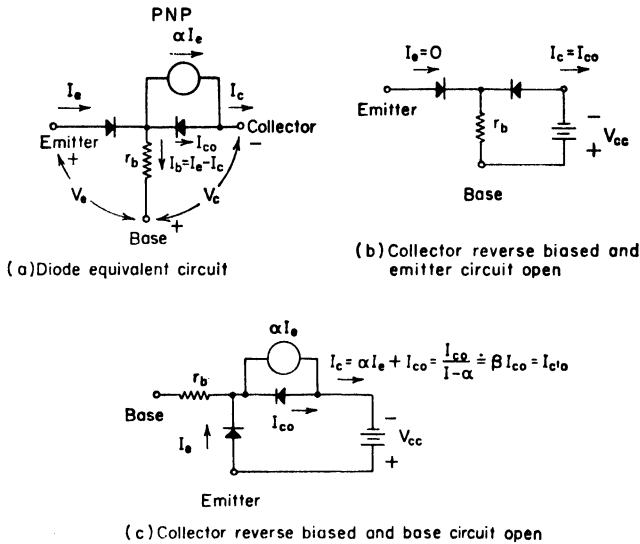



FIG. 4. Diode equivalent circuits.

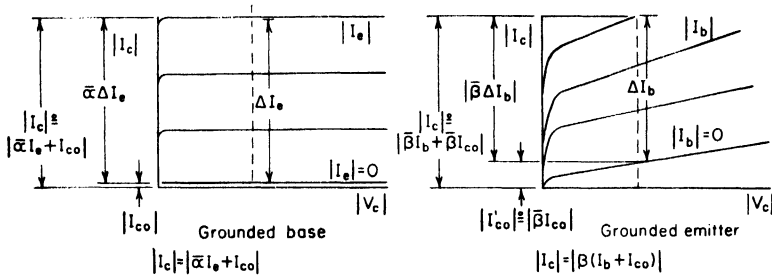
consist entirely of minority carriers injected into the base region and these carriers would diffuse without recombination to the collector and α would be unity. Actually the emitter is not completely efficient and recombination does take place; so α is less than unity and the difference between emitter and collector current appears in the base lead.

Power gain in the common-base configuration is possible because emitter current injected at the low-impedance level of the emitter circuit appears only slightly diminished at the collector terminal at a high-impedance level. In the common-emitter configuration current gain $\alpha(1 - \alpha)$ as well as power gain is provided. The common-emitter configuration is used most frequently, particularly in low-frequency circuits, because of its higher power gain, current gain greater than unity, and more favorable impedance levels for many applications.

The diode equivalent circuit is valuable in representing the transistor over a wide range of operating conditions and is particularly useful in considering large-signal operation. By considering the diodes to be nonideal junction types having finite resistance, saturation current when reverse biased, and a resistance inversely propor-

tional to current when forward biased, a fairly accurate representation of the transistor is obtained.

Figure 4*b* shows the diode equivalent circuit with the collector biased in the reverse direction and the emitter open-circuited. The current flowing in this connection, termed I_{co}, is made up of the saturation current of the collector diode plus current due to leakage. For reverse biases of more than a few tenths of a volt the saturation current is constant, and for low voltages or low leakage, I_{co} is made up largely of the saturation current. The saturation current is strongly temperature-dependent, approximately doubling for each 10°C increase in temperature. Typical values of I_{co} are 5 to 20 μa for low-power germanium junction transistors at 25°C. For silicon transistors it is much lower, typically of the order of 0.1 μa at the same temperature. The variation of I_{co} is one of the major difficulties in operating transistors over wide temperature limits and is responsible for the possibility of a thermal runaway condition wherein a rise in temperature increases I_{co}, causing an increase in power dissipation resulting in further rise in temperature and I_{co} until the transistor is ruined.

Figure 4*c* shows the diode equivalent circuit with the collector diode reverse biased and the base circuit open. The collector current in this connection is termed I_{co}' and since I_e must equal I_c, I_{co}' must be $I_{co}/(I - \alpha)$, which is approximately equal to $\beta I_{co}[\beta = \alpha/(1 - \alpha)]$. I_{co}' can be very much greater than I_{co} because beta may range from 20 to 100 or more. The condition illustrated in Fig. 4*c* is the worst situation as far as collector current variations with temperature and I_{co} are concerned while that of Fig. 4*b* is the best. In actual operation, with typical bias arrangements, conditions are intermediate between those of Figs. 4*b* and 4*c*.

Where:

| | Designates magnitudes only, thus eliminating signs Therefore this information can be used for both PNP & NPN type transistors.

$\bar{\beta} = \frac{\bar{\alpha}}{1-\bar{\alpha}} \cong \frac{1}{1-\alpha}$ when value of $\bar{\alpha}$ is close to 1

$\bar{\beta}$ = D-C value of β

$\bar{\alpha}$ = D-C value of α

FIG. 5. Collector output characteristics for junction transistors, d-c conditions.

11. Characteristic Curves. Figure 5 shows the curves of collector I_c current vs. collector voltage with emitter and base currents as parameters and showing the effects of I_{co}. Also given are the important d-c relationships. Figure 6 shows actual collector characteristics for a typical germanium-junction transistor. In Fig. 6*a*, for the common-base connection, note that the curves do not break until the collector-base voltage is reversed, implying that the transistor may be operated at very low collector-to-base voltages. The slope of the common-base curves represents a much higher resistance than for the common-emitter circuit and the current gain for the common-emitter circuit $\left(\beta = \left.\frac{\Delta I_c}{\Delta I_b}\right|_{V_c} = \frac{\alpha}{1 - \alpha}\right)$ is much greater than the current gain for the common-base circuit $\left(\alpha = \left.\frac{\Delta I_c}{\Delta I_e}\right|_{V_c}\right)$.

(*a*)

(*b*)

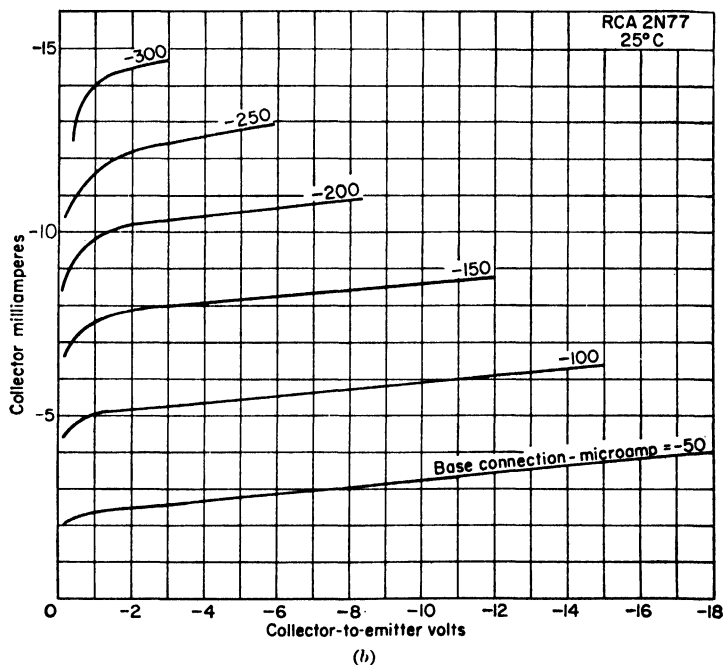
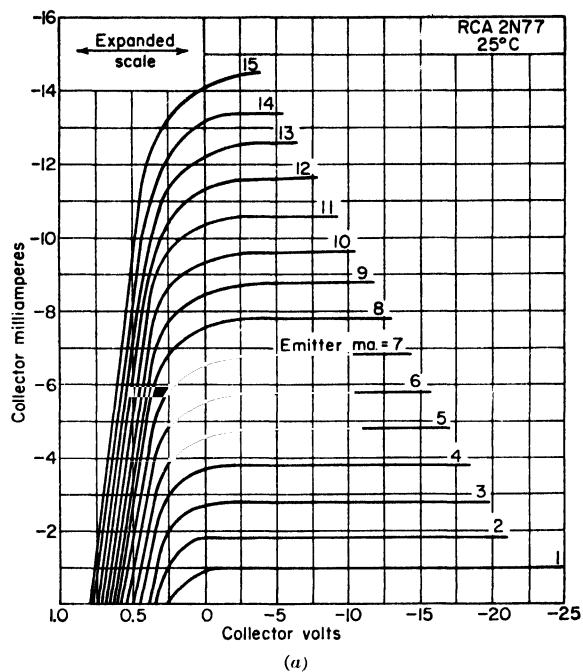



FIG. 6. Collector characteristics of typical germanium-junction transistor, RCA 2N77.

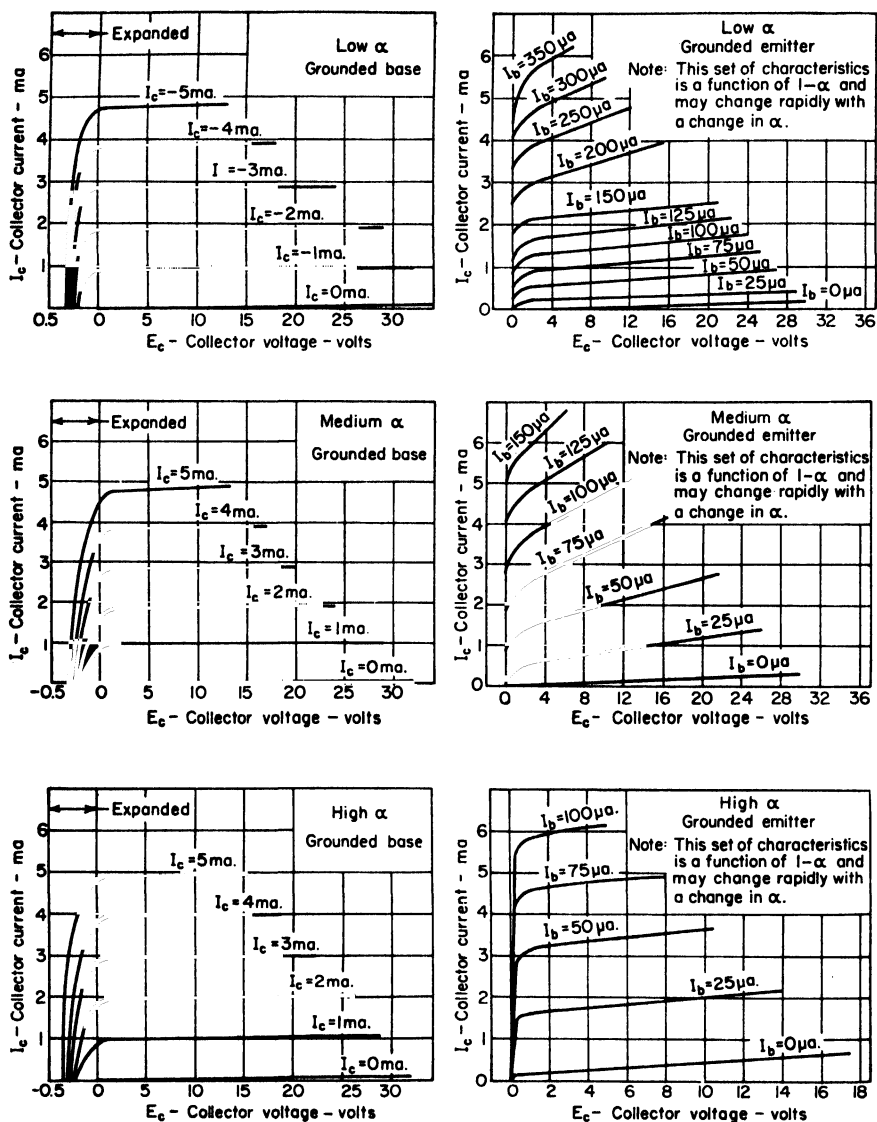



FIG. 6c. Output characteristic curves for *n-p-n* low-power, low-frequency grown-junction triode transistors. (*Texas Instruments, Inc.*)

Figure 6*c* shows a family of output curves for both the grounded-base and grounded-emitter circuits for low-, medium-, and high-alpha conditions.

Examination of Fig. 6 shows that collector voltage is a function of both collector current and emitter current (or base current). Similarly, emitter voltage is a function of emitter current and collector current. That is,

$$V_e = F_1(I_c, I_e)$$

and

$$V_c = F_2(I_c, I_e)$$

when currents and voltages are total values. For small variations in the independent variables,

$$\Delta V_e = \left.\frac{\Delta V_e}{\Delta I_e}\right|_{I_c=\text{const}} \times \Delta I_e + \left.\frac{\Delta V_e}{\Delta I_c}\right|_{I_e=\text{const}} \times \Delta I_c$$

$$\Delta V_c = \left.\frac{\Delta V_c}{\Delta I_e}\right|_{I_c=\text{const}} \times I_e + \left.\frac{\Delta V_c}{\Delta I_c}\right|_{I_e=\text{const}} \times \Delta I_c$$

where the Δs indicate small changes. The ratios of changes in voltage to changes in current may be identified as incremental resistances and changes in voltages and currents as small-signal variations about the bias values. Then the above equations may be written as

$$V_e = r_{11}I_e + r_{12}I_c$$
$$V_c = r_{21}I_e + r_{22}I_c$$

where the voltages and currents are now small-signal values. A small-signal common-base equivalent circuit for the transistor may be made using the r parameters as given above and shown in Fig. 7.

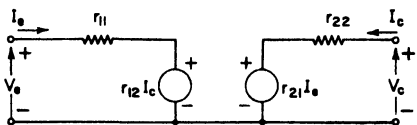



FIG. 7. Two-generator small-signal common-base equivalent circuit.

12. Impedance, Hybrid, and Admittance Parameters. The low-frequency r parameters may be determined from the characteristic curves of the transistors. R_{11} is the slope of the V_eI_e curve for a constant collector current, r_{12} is the ratio of the change in emitter voltage for a small change in collector current for a constant value of emitter current, r_{21} is the ratio of the change in collector voltage for a small change in emitter current for a constant value of collector current, and r_{22} is the slope of the V_cI_c curves for constant emitter current. The two-generator equivalent circuit of Fig. 7 is not too convenient and the T equivalent of Fig. 8a is most frequently used.

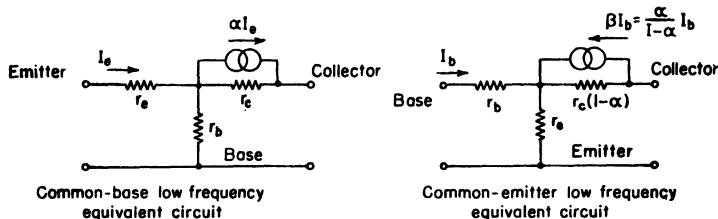
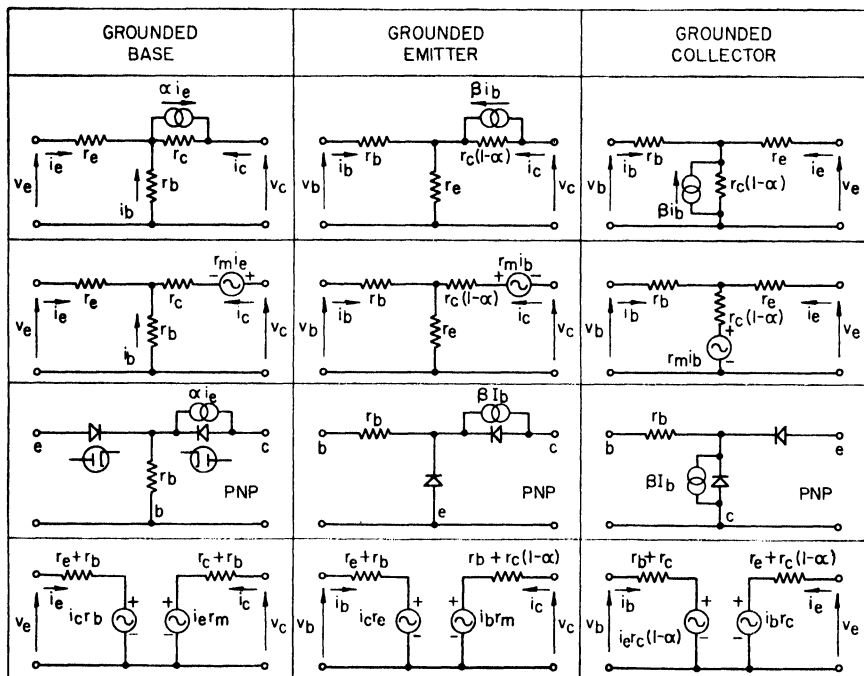
Table 3. Relation between T-equivalent Parameters and r and h Parameters

Parameters of T-equivalent circuit	Common-base parameters		Common-emitter parameters	
	r	h	r	h
α	$\approx \frac{r_{21}}{r_{22}}$	$-h_{21}$		
$\beta\left(=\frac{\alpha}{1-\alpha}\right)$			$\approx \frac{r_{21}}{r_{22}}$	h_{21}
r_e	$r_{11} - r_{12}$	$h_{11} - \frac{h_{12}(1+h_{21})}{h_{22}}$	r_{12}	$\frac{h_{12}}{h_{22}}$
r_b	r_{12}	$\frac{h_{12}}{h_{22}}$	$r_{11} - r_{12}$	$h_{11} - \frac{h_{12}(1+h_{21})}{h_{22}}$
r_c	$r_{22} - r_{12}$	$\frac{1}{h_{22}}$	$r_{22} - r_{21}$	$\frac{1+h_{21}}{h_{22}}$

Other equivalents are shown in Fig. 8b for the different transistor configurations, and Table 3 is useful for determining the elements of the T equivalent from the r and h parameters for the common-base and common-emitter circuits. Table 4 gives the relationships between the r, h, and y parameters to the elements of the T-equivalent circuit.

The common-base h parameters are arrived at in a process similar to that used in obtaining the r parameters by considering collector voltage and emitter current as independent variables. The h parameters are particularly useful in making measurements on junction transistors because of the high impedance of the collector circuit and low impedance of the emitter circuit under normal bias conditions.

(a)

GROUNDED BASE	GROUNDED EMITTER	GROUNDED COLLECTOR

(Note: direction of current flow is that of conventional network theory)

(b)

FIG. 8. Small-signal low-frequency T-equivalent circuits.

The common-base y parameters are obtained by considering the emitter voltage and collector voltage as independent variables. The common-emitter parameters are obtained using base voltage and base current as variables instead of emitter voltage and current.

The transistor parameters are functions of bias conditions and temperature, as illustrated by the curves of Fig. 9 for a typical silicon audio transistor.

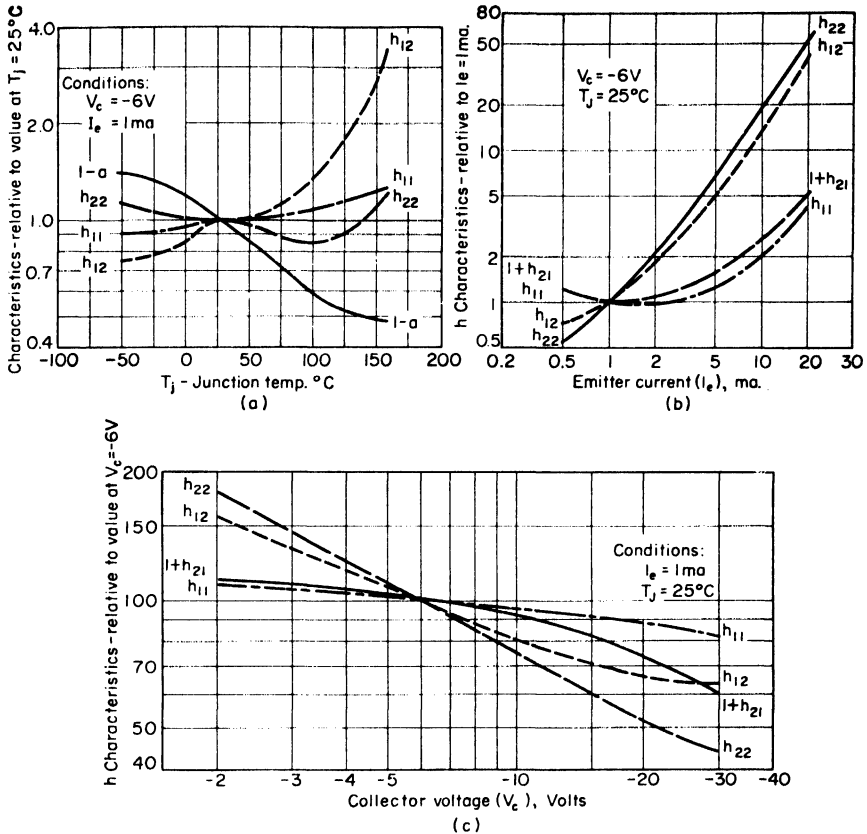



FIG. 9. Variation of transistor parameters with bias conditions and temperature (Raytheon CK790).

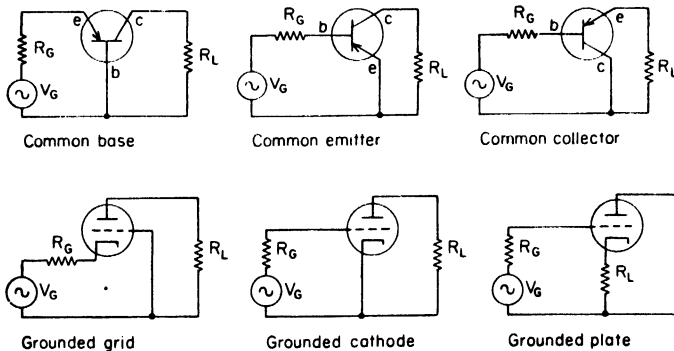



FIG. 10. Basic transistor configurations and analogous tube circuits.

TRANSISTOR APPLICATIONS

18. Low-frequency Amplifiers. The three basic circuits for transistor amplifiers are illustrated in Fig. 10 together with their vacuum-tube analogies. Their characteristics relative to one another are shown in Sec. 7.

Table 5 gives approximate expressions for current, voltage and power gain, input and output resistances, and optimum values of load resistance for the three transistor configurations. Figure 11 shows the variation of power gain and input resistance as

Table 4

Conversion Formulas for Impedance Parameters

Function	Grounded base	Grounded emitter	Grounded collector
Input impedance, ohms $R_{11} = \frac{v_1}{i_1}\Big\|_{i_2=0}$	$r_e + r_b$	$r_e + r_b$	r_c
Output impedance, ohms $R_{22} = \frac{v_2}{i_2}\Big\|_{i_1=0}$	r_c	$r_c(1 - \alpha)$	$r_c(1 - \alpha)$
Backward transfer impedance, ohms $R_{12} = \frac{v_1}{i_2}\Big\|_{i_1=0}$	r_b	r_e	$r_c(1 - \alpha)$
Forward transfer impedance, ohms $R_{21} = \frac{v_2}{i_1}\Big\|_{i_2=0}$	αr_c	$-\alpha r_c$	r_c

Assume $r_c \gg r_b$, $r_c \gg r_e$.
Above open-circuit parameters derived from equations $v_1 = i_1 r_{11} + i_2 r_{12}$ and $v_2 = i_1 r_{21} + i_2 r_{22}$.

Conversion Formulas for Hybrid Parameters

Function	Grounded base	Grounded emitter	Grounded collector
Input impedance, ohms $H_{11} = \frac{v_1}{i_1}\Big\|_{v_2=0}$	$r_e + r_b(1 - \alpha)$	$r_b + \frac{r_e}{1 - \alpha}$	$r_b + \frac{r_e}{1 - \alpha}$
Output admittance, mhos $H_{22} = \frac{i_2}{v_2}\Big\|_{i_1=0}$	$\frac{1}{r_c}$	$\frac{1}{r_c(1 - \alpha)}$	$\frac{1}{r_c(1 - \alpha)}$
Voltage feedback ratio, no units $H_{12} = \frac{v_1}{v_2}\Big\|_{i_1=0}$	$\frac{r_b}{r_c}$	$\frac{r_e}{r_c(1 - \alpha)}$	1
Current gain, no units $H_{21} = \frac{i_2}{i_1}\Big\|_{v_2=0}$	$-\alpha$	β	$\frac{-1}{1 - \alpha} \doteq \beta$

Assume $r_c \gg r_b, r_e$.
Above H parameters derived from equations $v_1 = H_{11}i_1 + H_{12}v_2$ and $i_2 = H_{21}i_1 + H_{22}v_2$.

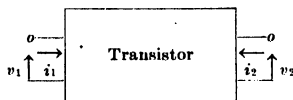



Table 4 (*Continued*)

Conversion Formulas for Admittance Parameters

Function	Grounded base	Grounded emitter	Grounded collector
Input admittance, mhos $Y_{11} = \left.\frac{i_1}{v_1}\right\vert_{v_2=0}$	$\frac{1}{r_b(1-\alpha)+r_e}$	$\frac{1}{r_b + \frac{r_e}{1-\alpha}}$	$\frac{1}{r_b + \frac{r_e}{1-\alpha}}$
Output admittance, mhos $Y_{22} = \left.\frac{i_2}{v_2}\right\vert_{v_1=0}$	$\frac{r_b + r_e}{r_c[r_b(1-\alpha)+r_e]}$	$\frac{1}{\frac{\alpha r_c r_e}{r_e + r_b} + r_c(1-\alpha)}$	$\frac{1}{r_e + r_b(1-\alpha)}$
Backward transfer admittance, mhos $Y_{12} = \left.\frac{i_1}{v_2}\right\vert_{v_1=0}$	$\frac{-r_b}{r_c[r_b(1-\alpha)+r_e]}$	$\frac{-1}{r_e + r_b\left[1 + \frac{r_c(1-\alpha)}{r_e}\right]}$	$\frac{-1}{\frac{r_c}{1-\alpha} + r_b}$
Forward transfer admittance, mhos $Y_{21} = \left.\frac{i_2}{v_1}\right\vert_{v_2=0}$	$\frac{-\alpha}{r_b(1-\alpha)+r_e}$	$\frac{\alpha}{r_b(1-\alpha)+r_e}$	$\frac{-1}{r_e + r_b(1-\alpha)}$

Above short-circuit parameters derived from equations $i_1 = v_1y_{11} + v_2y_{12}$ and $i_2 = v_1y_{12} + v_2y_{22}$.

Table 5. Small-signal Low-frequency Audio Amplifier Formulas

Function	Grounded base	Grounded emitter	Grounded collector
Current amplification A	$-\alpha$	β	β
Voltage amplification V_i	$\frac{R_L}{\frac{r_e}{\alpha} + \frac{r_b}{\beta}}$	$\frac{-R_L}{\frac{r_e}{\alpha} + \frac{r_b}{\beta}}$	1
Power gain G_p	$\alpha^2 \times \frac{R_L}{R_{in}}$	$\beta^2 \times \frac{R_L}{R_{in}}$	β
Input resistance R_{in}	$r_e + r_b(1-\alpha)$	$r_b + (\beta \times r_e)$	$R_L \times \beta$
Output resistance R_o	$\left(\frac{\frac{r_b}{\beta} + R_g}{r_b + R_g}\right) \times r_c$	$\frac{r_c}{\beta} + \left(r_e \times \frac{R_g + r_m}{R_g + r_b}\right)$	$r_e + \frac{r_b + R_g}{\beta}$
Optimum value of load resistance R_L	$\frac{r_c}{4}$	$\frac{2r_c}{\beta}$	$\frac{R_{in}}{50}$ $(500 < R_L < 2{,}500)$
Under the following conditions	$\beta + 1 \doteq \beta$ $R_L \ll r_c$ Value of α close to 1 $\beta = \frac{\alpha}{1-\alpha} \doteq \frac{1}{1-\alpha}$	$\beta + 1 \doteq \beta$ $R_L \ll r_c(1-\alpha)$ Value of α close to 1 $\beta = \frac{\alpha}{1-\alpha} \doteq \frac{1}{1-\alpha}$	$\beta + 1 \doteq \beta$ $R_L \ll r_c(1-\alpha)$ $R_g \ll r_c$ Value of α close to 1 $\beta = \frac{\alpha}{1-\alpha} \doteq \frac{1}{1-\alpha}$

a function of load resistance for a typical audio transistor for the three configurations. The maximum for power gain is quite broad, indicating that the load resistance for maximum gain is not critical. The input resistance increases with increasing load resistance for the common-base circuit, whereas the opposite is true for the common-emitter circuit. In both instances the input resistance is increased and the gain

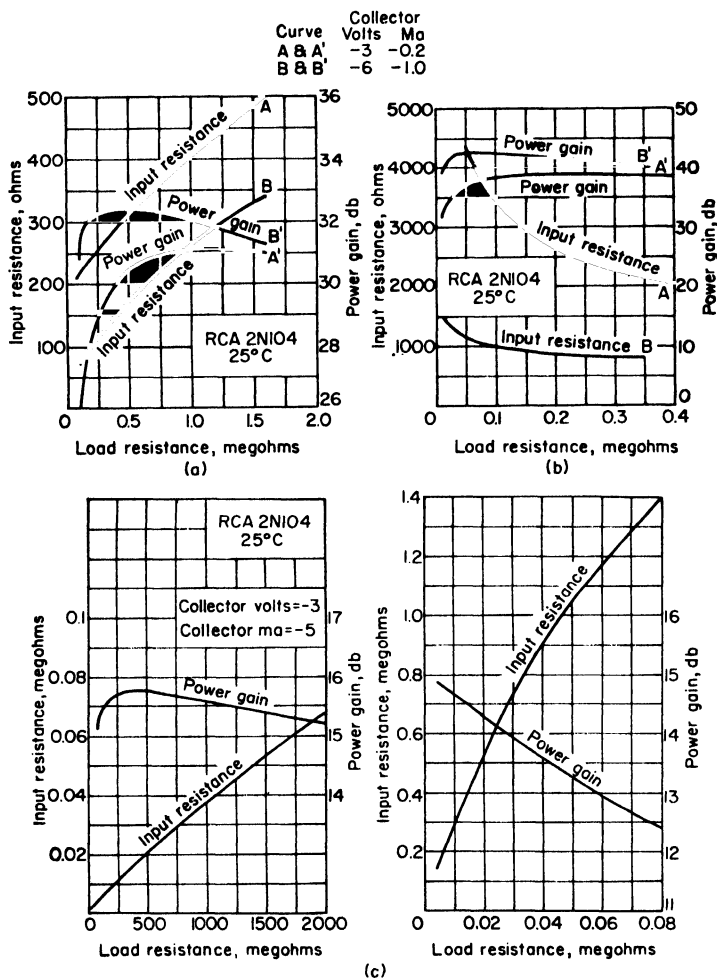



FIG. 11. Gain and input resistance of typical audio junction transistor. (*a*) Common base; (*b*) common emitter; (*c*) common collector.

decreased as the emitter current is decreased. The input resistance for the common-collector circuit is almost directly proportional to the load resistance [approximately $R_L/(1 - \alpha)$] and high input resistances may be achieved.

14. Temperature Effects. The stability of the operating point of a transistor amplifier with changes in I_{co} due to temperature changes and in replacing transistors is an important consideration in biasing transistors. The magnitudes of the effect of these changes are a function of the biasing circuits. The transistor is normally biased

at the mid-point of the collector load line, but changes in temperature or replacement with another transistor will cause the operating point to change and may result in clipping of a portion of the signal or in extreme cases may result in the transistor's being saturated or cut off. Figure 12 represents the d-c circuit of the transistor (any configuration) with supply voltages shorted. R_b represents the d-c resistance seen in the base circuit (Thévenin equivalent resistance) plus the internal base resistance.

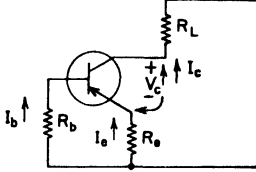



FIG. 12. Bias source resistance.

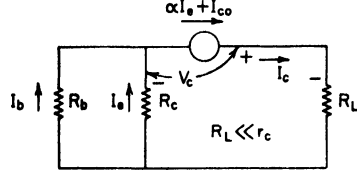



FIG. 13. Circuit for determining effects of I_{co} on operation for equivalent of Fig. 12.

R_e is the resistance seen in the emitter circuit plus the internal emitter resistance; R_L is the d-c resistance seen in the collector circuit. Figure 13 is an equivalent circuit of Fig. 12 with the assumption that the collector resistance is much larger than R_L, which is usually the case.

The factors which relate changes in bias conditions with changes in I_{co} may be obtained from Fig. 12, where it is seen that

$$I_b = \frac{R_e}{R_b} I_e \qquad I_b + I_e = I_c$$

(using current directions shown in Fig. 12) and $I_c = \alpha I_e + I_{co}$. Using these relationships, the expression $I_e = \dfrac{I_{co}}{(1 - \alpha) + R_e/R_b}$ is obtained which may be differentiated with respect to I_{co} to obtain the stability factor for emitter current:

$$S_e = \frac{\delta I_e}{\delta I_{co}} = \frac{1}{(1 - \alpha) + R_e/R_b} \quad \text{or} \quad \frac{R_b}{(1 - \alpha)R_b + R_e}$$

S_e gives the change in emitter current for a change in I_{co} for a particular biasing arrangement assuming that other parameters have not changed (for example, if S_e is 10 and I_{co} changes by 10 μa because of a change in temperature or substitution of a similar transistor except for a difference in I_{co}, then the emitter current would change by 100 μa).

Differentiating the expression $I_c = \alpha I_e + I_{co}$ with respect to I_{co} gives the collector-current stability factor

$$S_c = \frac{\delta I_c}{\delta I_{co}} = \alpha \frac{\delta I_e}{\delta I_{co}} + 1 = \alpha S_e + 1$$

or

$$S_c = \frac{R_b + R_e}{(1 - \alpha)R_b + R_e}$$

In a similar manner the stability factor for collector voltage V_c may be obtained and is

$$S_v = -(R_L S_c + R_e S_e) = \frac{-(R_b R_L + R_b R_e + R_e R_L)}{(1 + \alpha)R_b + R_e}$$

The stability factor should be low to minimize variations in operating point with temperature or in replacing transistors. Examination of the expression for S_c shows that, for large R_b (constant base current bias) or for small R_e, S_c approaches $1/(1 - \alpha)$, which may be very large (of the order of 50), and for small R_b or large R_e approaches 1. The stability of the operating point is thus improved by keeping R_b, the d-c resistance in the base circuit, as small as possible and R_e, the d-c resistance in the emitter circuit, as high as possible. This is most easily accomplished in transformer-coupled circuits. It is more difficult with RC-coupled circuits because, in the common-emitter and common-collector circuits, the base bias network shunts the input to the

transistor. Making S small requires more d-c power; so a compromise is made between stability of operating point and power dissipation in the bias network. Typical biasing methods are shown in circuit diagrams of various amplifier circuits to follow.

For circuits having feedback resistances connected between transistor terminals the stability factors should be calculated for the particular circuit.

15. Transformer Coupling. The transistors may be cascaded in amplifier circuits using transformer, impedance, RC, or direct coupling or combinations of these. With transformer coupling gain may be achieved with all three configurations with the common-emitter circuit providing the most gain and the common-collector circuit the least. Transformer coupling has the advantages of providing an impedance match between output of one stage and input of the next stage leading to maximum power

(a) Common-base circuit with two battery supply

(b) Common-base circuit with single battery supply

(c) Common-emitter circuit with two battery supply

(d) Common-emitter circuit with single battery supply

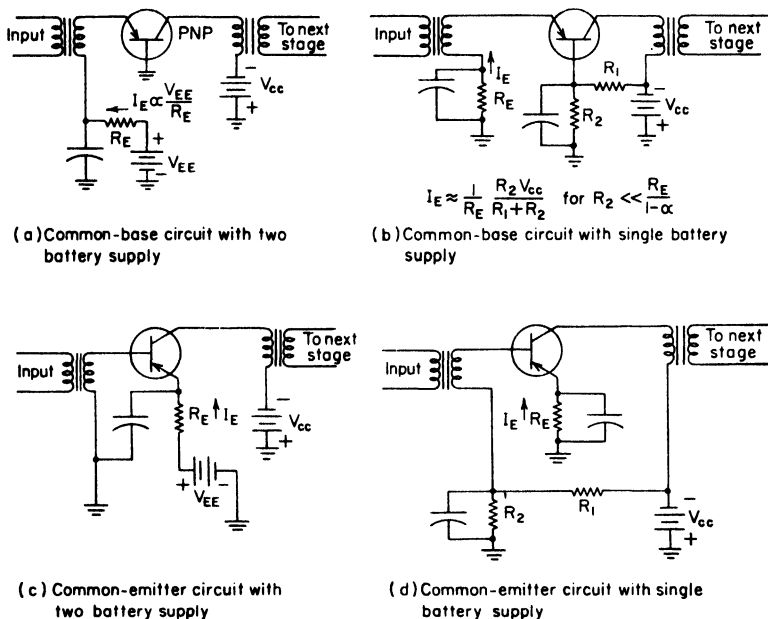



FIG. 14. Transformer-coupled transistor circuits.

gain, simplifying the problem of stabilizing the transistor against changes in the operating point with temperature and transistor parameter variations, and providing high efficiency. The disadvantages include the expense, weight, and bulk and the limited frequency response of transformers. Because of the high output impedances of the common-base and common-collector circuits it is not always practical to match impedances since this may lead to a situation where the maximum collector voltage is exceeded at the output voltage peaks. The impedance-coupled circuit has the advantage of high efficiency operation but does not permit matching impedances and therefore cannot be used in the common-base configuration. Figure 14 shows typical transformer-coupled circuits and bias arrangements.

16. RC Coupling. In resistance-capacitance coupled circuits only common-emitter and common-collector configurations provide gain when connected in cascade, the common-emitter circuit being most widely used in intermediate stages because of its higher gain (typically of the order of 30 db). The maximum power gain may not be obtained in RC-coupled circuits because input and output impedances cannot be

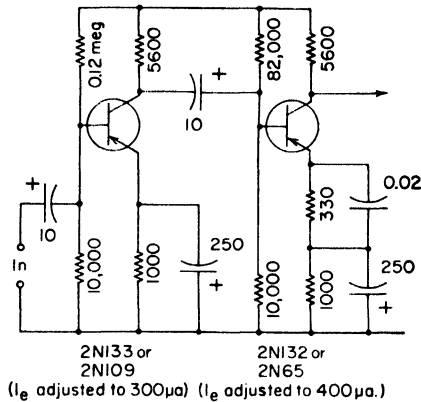
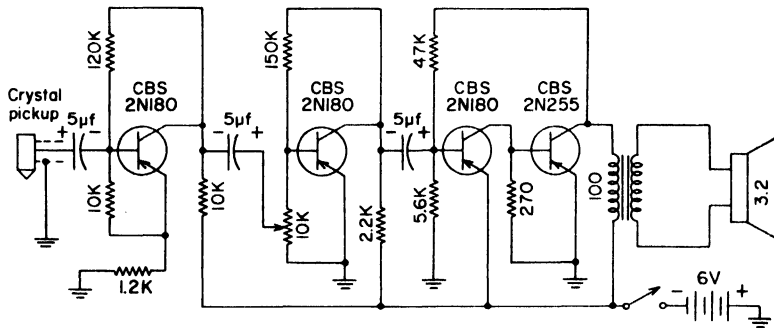
matched but the loss is not excessive and is compensated by the circuit being less expensive, lighter, and more compact. The choice of input stage depends upon the input impedance desired: for low input impedance the common-base configuration is indicated (provides impedance match but no gain); for intermediate values of input impedance, the common-emitter circuit and, for high values of input impedance, the common-collector stage are best. In the output stage the common-collector provides a low-impedance output, the common-emitter an intermediate value of output impedance, and the common-base circuit a high output impedance. The power gain of a multistage amplifier (ratio of output power to input power) is given by $A_i^2R_L/R_{in}$,

(a) Microphone preamplifier

(b) Portable phonograph amplifier

FIG. 15. Circuits illustrating RC coupling.

where R_L is the output load resistance, R_{in} is the input resistance of the first stage, and A_i is the product of the current gains of the individual amplifier stages.

Circuits illustrating RC coupling are shown in Fig. 15. Figure 15a shows an amplifier having 49 db gain from a 500-ohm source to a 5,600-ohm load and a frequency response flat within 0.5 db from 20 to 20,000 cps. The 0.02-μf capacitor in shunt with the 330-ohm resistor in the emitter circuit is to compensate the drop in gain of the transistor at high frequencies.[1]

In the circuit of Fig. 15b a high input impedance is obtained by placing an unbypassed resistor in the emitter circuit [appears at the base circuit as $R_e/(1 - \alpha)$] of the

[1] Stark, H. F., Transistor Preamps, *Audio*, April, 1956.

first stage and returning the 10,000-ohm bias resistor to the emitter. The bias stability is improved, at the expense of some degeneration, by connecting the base bias divider to the collector.

In RC amplifiers the low-frequency cutoff is determined by the coupling capacitor and input and output resistances of the coupled stages (assuming that bias resistors in the emitter circuit, etc., are adequately by-passed). The frequency at which the response is down 3 db from the mid-band value is given by

$$F_L = \frac{1}{2\pi C(R_o + R_i)}$$

where F_L is the low-frequency 3-db point, C is the capacitance of the coupling capacitor, R_o is the output resistance of the driving stage (essentially the collector load resistance, which is much smaller than the transistor output resistance in most instances), and R_i is the input resistance of the driven stage. The high-frequency cutoff is determined for the most part by the cutoff frequency of the transistor, f_β for common-emitter and common-collector stages and f_α for common-base stages $[f_\beta = f_\alpha/(1 - \alpha_0)]$, where α_0 is the low-frequency α.

17. Direct Coupling. Transistors may be coupled directly to provide gain down to d.c. or to minimize the number of components. The principal difficulty with

(a) TR₃ is for temperature compensation (b) D is germanium diode for temperature compensation (c) $r_{(T)}$ is ceramic thermistor for temperature compensation

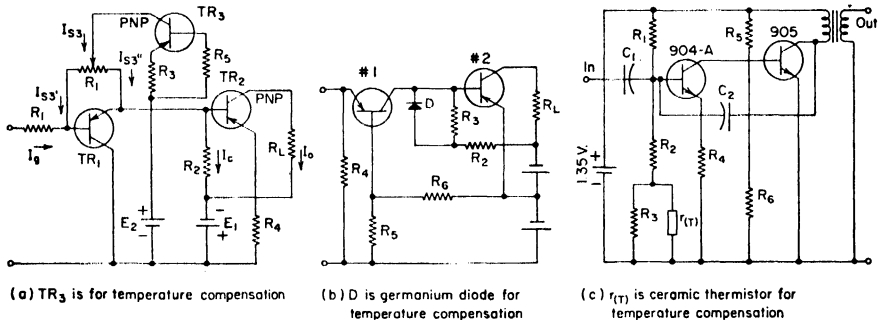



FIG. 16. Amplifiers employing direct coupling.

direct-coupled circuits is the drift in operating point with changes in temperature or transistor parameters, particularly in the first stages of a multistage circuit since bias changes in these stages appear amplified in the succeeding ones. One method of compensation is to use configurations in which changes in I_{co} with temperature tend to balance out (see Table 6). For a two-stage amplifier the condition for this type of stabilization is

$$A_{i2}S_1i_{co1} + S_2i_{co2} = 0$$

where A_{i2} is the current gain of the first stage, S is the stability factor (ratio of change in output current to a change in I_{co}), and i_{co} is the change in I_{co} and the number subscripts refer to stage number. This condition leads to the permissible arrangements shown in Fig. 16 for this type of temperature compensation. This method is not entirely practical since it involves sorting transistors according to I_{co} and placing them in the proper stage. Other means for temperature compensation are to use temperature-sensitive elements such as transistors, semiconductor diodes, thermistors, or other temperature-sensitive resistances. Some practical amplifiers having direct coupling are shown in Fig. 16, and in Fig. 15*b* the driven and output stage are direct-coupled.[1] The problem of I_{co} variations is minimized by use of silicon transistors in preference to germanium types since I_{co} is much lower. Elaborate means of temperature com-

[1] Stanton, J. W., A Transistorized D-C Amplifier, *Trans. IRE*, **PGCT**, 65, March, 1956 (Fig. 16*a*). Keonjian, Edward, Temperature Compensated D-C Transistor Amplifiers, *Proc. IRE*, April, 1954 (Fig. 16*b*), p. 668; Micropower Audio Amplifier, *Trans. IRE*, Circuit Theory, March, 1956 (Fig. 16*c*), p. 68.

pensation of course is not necessary in environments where the range of temperature variation is not great or where changes in bias conditions due to I_{co} variations are not significant compared with normal bias.

18. Audio Output Power Amplifiers. Power-amplifier stages differ considerably from the small-signal circuits previously described. With small-signal circuits the object is to obtain maximum gain; with power amplifiers the object is to obtain the maximum output power at the highest efficiency subject to the constraints of maximum collector voltage, maximum collector current, power dissipation, and operating temperature placed upon the transistor. In obtaining high power output, the load is much different from the output impedance of the transistor circuit and no attempt is made to provide an impedance match to maximize gain; *i.e.*, because of the constraints mentioned above, loads providing maximum power gain are not consistent with loads giving maximum power output. In power amplifiers, the large signals employed result in variation of the transistor parameters with the signal, thus causing distortion. Large-signal or average values rather than small-signal parameters should be used.

Fig. 17. Output characteristics showing limits of operation.

Transistors having permissible power dissipations as high as about 20 watts and maximum collector currents of up to about 4 amp are available for high-power-output stages.

Figure 17 is a sketch of the output characteristics of a transistor with the limits imposed by maximum power dissipation, maximum collector current, and maximum collector voltage shown. When the area of permissible operation is specified by the

Table 6. Circuit Arrangements in Which I_{co} Variations with Temperature Tend to Compensate

For transistors of same type (*p-n-p* or *n-p-n*)		For transistors of dissimilar types (*p-n-p* and *n-p-n* or *n-p-n* and *p-n-p*)	
First stage	Second stage	First stage	Second stage
Common-emitter	Common-emitter	Common-base	Common-base
Common-base	Common-emitter	Common-emitter	Common-base
Common-base	Common-collector	Common-collector	Common-emitter
Common-emitter	Common-collector	Common-collector	Common-collector
Common-collector	Common-base		

maximum power dissipation, collector voltage, and collector current, the load R_L giving the maximum output is determined by two of these quantities. In Fig. 17 two situations are illustrated by showing the load lines resulting for two values of maximum collector current, $I_{c\,max}$ and $I_{c'\,max}$. The maximum power output is obtained when the collector voltage and current are permitted to reach their maximum values.

This leads to load lines R_L and R_L' for the two cases. Load line R_L is permissible since it falls within the area where the maximum limits of the transistor are not exceeded; load line R_L', on the other hand, would not be permissible since the maximum power dissipation would be exceeded. Therefore, if $I_{c' \max}$ is specified as the maximum collector current as shown in Fig. 17, then the load line for maximum output power is dictated by the maximum power dissipation and either $V_{c \max}$ or $I_{c' \max}$. Because of the dropoff of current gain with high collector currents it is best to operate at low values of collector current. For the situation where maximum power dissipation is a limiting factor a load line R_L'' drawn from $V_{c \max}$ and tangent to the maximum power-dissipation hyperbola is the solution.

19. Power Output. In single-transistor class A amplifiers using a transformer in the output circuit, obtaining maximum power output (peak collector voltage of $V_{c \max}$) means using a supply voltage of $V_{c \max}/2$. If some other supply voltage less than $V_{c \max}/2$ is used, then the maximum output power will be limited by the maximum permissible power dissipation of the transistor or by the maximum collector current.

FIG. 18. Mounting details, Minneapolis-Honeywell power transistor.

If it is assumed that the transistor is driven along the load line from zero collector voltage to zero collector current (which is nearly the case with transistors) then the maximum output power obtainable is

$$P_{0 \max} = \frac{V_{c \max} I_{c \max}}{8}$$

for situations where output power is restricted by the maximum collector voltage and maximum collector current. The load resistance R_L is $V_{c \max}/I_{c \max}$.

The maximum power output, for the situation where the output power is limited by the maximum collector voltage and the maximum power dissipation, is $P_{0 \max} = \frac{1}{2}P_d$ and the load resistance is $R_L = V_c{}^2{}_{\max}/4P_d$, where P_d is the allowable power dissipation. The power output in general, for class A amplifiers with a transformer in the output, in terms of the supply voltage V_{cc} is $P_0 = V_{cc}^2/2R_L$ and the collector voltage swings from zero to $2V_{cc}$ at the peaks. The efficiency of the class A circuit approaches a maximum theoretical value of 50 per cent for an ideal amplifier.

Output power is limited by the maximum allowable collector voltage, current, and power dissipation of the transistor. It is not dependent on matching output impedances and is essentially the same for common-emitter, common-base, and common-collector configurations. The allowable power dissipation is a function of the junction temperature, and the transistor must be derated as the temperature increases in accordance with manufacturer's instructions. Transistors designed specifically for power applications are generally constructed to permit mounting in good thermal contact with a heat sink (such as a metal chassis), as shown in Fig. 18.

20. Output Gain. Power gain, unlike power output, is dependent on the transistor configuration. Approximate expressions for power gain for the three basic transistor configurations are

For the common-base circuit, $$G_p \approx \frac{\alpha^2 R_L}{r_e + (1 - \alpha)r_b}$$

For the common-emitter circuit, $$G_p \approx \frac{[\alpha/(1 - \alpha)]^2 R_L}{r_b + r_e/(1 - \alpha)}$$

For the common-collector circuit, $$G_p \approx \frac{[1/(1 - \alpha)]^2 R_L}{r_b + (r_e + R_L)/(1 - \alpha)}$$

where it is assumed that r_e is much larger than R_L. Figure 19 is a plot of gain vs. load resistance for the three configurations for some assumed values of r_e, r_b, and α. The common-emitter circuit has the highest gain. The common-base circuit has a higher gain than the common-collector for high values of load resistance and lower gain for low values. The gain of the common-collector circuit is relatively independent of load resistance. The choice of circuit configuration is not to be made on the basis of gain alone since distortion is also to be considered. The emitter resistance is inversely proportional to emitter current, thus making the input impedance extremely nonlinear. Alpha drops off at high values of emitter current, causing further distortion. This effect is most bothersome in the common-emitter circuit, where the effect of the decreases in alpha is multiplied. Typical variations in current gain with collector current are shown in Fig. 20.[1] For low- or medium-power amplifiers where the reduction in alpha is not great over the operating range and where the load is sufficiently large to give good gain, the common-base circuit driven from a source of high impedance (to swamp out variations in input resistance) is capable of low distortion. The common-emitter circuit gives the best gain with the most distortion. The distortion in the common-emitter circuit may be reduced at a sacrifice in gain by the use

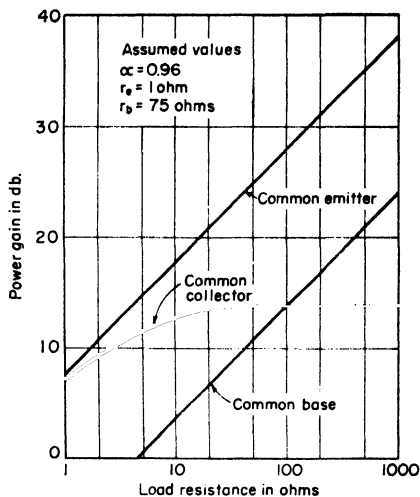



FIG. 19. Power gain vs. load resistance for the three basic configurations.

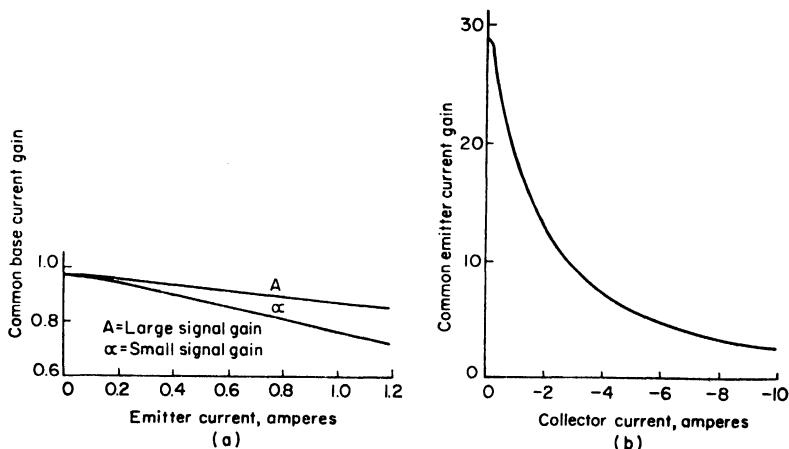



FIG. 20. Variation of common-base and common-emitter current with collector current, 70°F. (a) $V_c = 2$ volts; (b) $V_c = -2$ volts.

of feedback; however, the amount of feedback that may be used is limited, particularly in high-power transistors, by the phase shift of alpha and in the output trans-

[1] Mooers, H. T., "Design Procedures for Power Transistors," Minneapolis-Honeywell Technical Report TR 77C, Feb. 1, 1956.

former. For high power output with the associated low load resistance the gain of the common-collector circuit is higher than that of the common-base circuit and has lower distortion, particularly when driven by a low-impedance source. Thus for high power output with large collector-current variations, the common-collector configuration is the most satisfactory compromise between power gain and low distortion. Figure 21 shows some typical class A amplifiers. These circuits are biased at the mid-point of the load lines (also Fig. 15*b*).

21. Class B Amplifiers. Because the stand-by power in class A amplifiers is high and the efficiency limited to less than 50 per cent, class B circuits having low stand-by power and a maximum theoretical efficiency of 78.5 per cent are more suitable in many applications. The class B circuit utilizes transistors in a push-pull arrangement in which the transistors are biased at cutoff so that each transistor alternately amplifies opposite half cycles of the input signal. The same limitations as regards maximum collector voltage, current, and power dissipation of the transistor are to be observed. The collector supply voltage must be less than one-half of the maximum allowable collector voltage since the peak collector voltage reaches a value of twice the supply voltage. The maximum power dissipation should not be exceeded even for short periods with an average dissipation of less than maximum to avoid damage to the

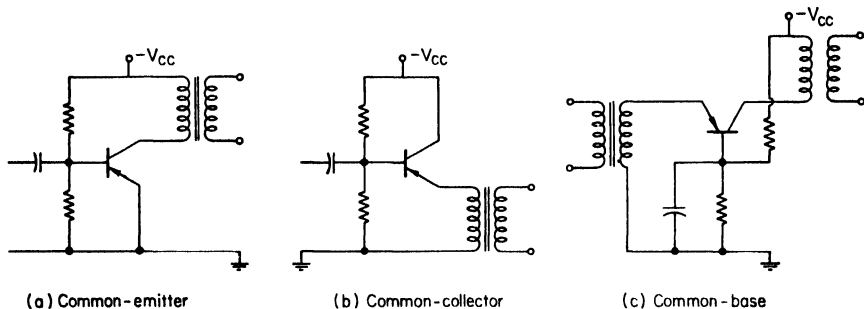



FIG. 21. Typical class A output circuits.

transistor. Strictly speaking, transistor audio amplifiers are not operated as pure class B for the transistors are slightly forward biased to minimize crossover distortion due to the nonlinear transfer characteristic at low signal levels where the transition between one transistor becoming active and the other becoming inactive occurs. This forward bias is quite small, amounting to about 0.1 volt emitter-to-base bias, and reduces the efficiency of the circuit slightly. Typical efficiencies of class B circuits are of the order of 60 to 70 per cent.

Output power in terms of the load resistance and the supply voltage is

$$P_0 = \frac{V_{cc}^2}{2R_L}$$

and the load resistance is $R_L = V_{cc}/I_{max}$, where I_{max} is the maximum collector current desired (at $V_c = 0$). (The collector-to-collector load at the terminals of the output transformer is $4R_L$.) The power dissipated by each transistor is approximately $0.069V_{cc}^2/R_L$ and the d-c power supplied to both transistors is about $0.64V_{cc}^2/R_L$.

For class B circuits the common-emitter and common-collector configurations are most popular, there being little to recommend the common-base circuit. The common-emitter circuit has a higher gain, thus requiring less driving power than the common-collector circuit, which has less distortion. The higher distortion of the common-emitter circuit may be reduced, with the previously mentioned difficulties, by the use of feedback or by use of nonlinear transfer characteristics. On the whole, however, the common-collector circuit appears to be the best compromise between

power gain and low distortion. Distortion in the common-collector circuit is particularly low when driven by a low-impedance source such as a common-collector driver. Another advantage of the common-collector configuration is that the output impedance is low, permitting a better low-frequency response with a given output transformer inductance or the same low-frequency response with a lower primary inductance as compared with the common-emitter circuit. The common-collector circuit requires a high-voltage drive (the output voltage follows the input voltage), thus necessitating a stepup transformer between the driver and the output stage or a higher supply voltage for the driver.

To minimize distortion the transistors should be fairly well matched in variation of current gain with collector current.

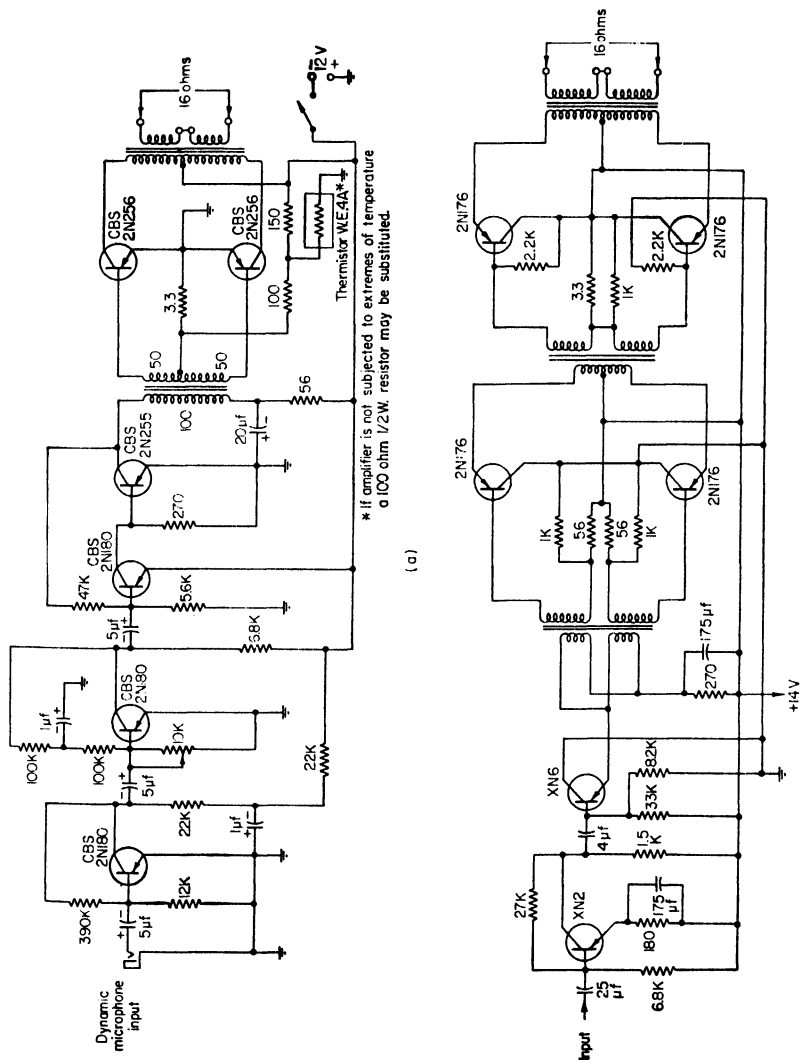



FIG. 22. Class B audio circuits. (*a*) 10-watt mobile public-address amplifier with class B common-emitter output. (*b*) 10-watt high-fidelity amplifier with common-collector output. (*c*) Appears on next page.

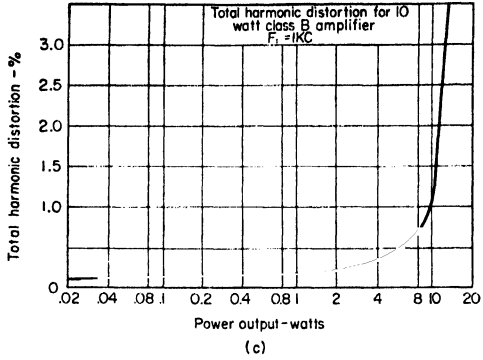



FIG. 22. (c) Distortion vs. output, amplifier of Fig. 22b.

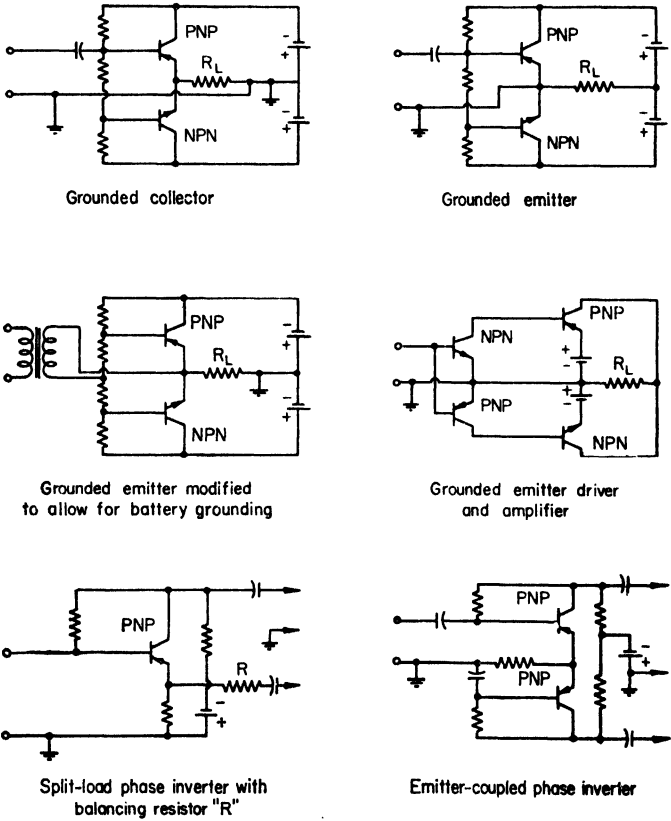



FIG. 23. Complementary symmetry class B circuits and phase-inverter circuits.

A novel class of circuits called *complementary symmetry* circuits results from the availability of transistors in *p-n-p* and *n-p-n* types. These permit push-pull output with single-ended input.

Figure 22 shows some class B audio circuits using transistors of a single polarity type.[1] Figure 23 shows some class B complementary symmetry circuits and some phase inverters for obtaining a push-pull driving source.

HIGH-FREQUENCY AMPLIFIERS

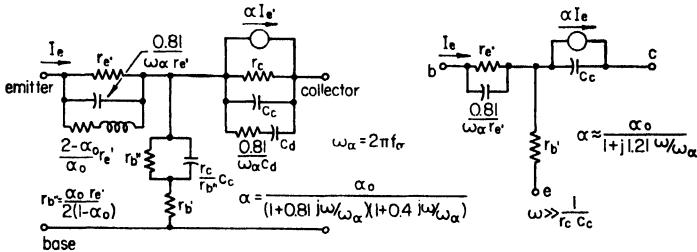
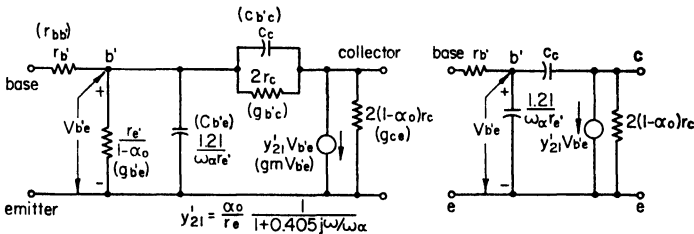
22. Equivalent Circuits. At higher frequencies the equivalent circuits of transistors become more complicated than in the middle range of frequencies where all the elements of the equivalent circuit must be retained. At these high frequencies the equivalent circuit can become simplified and satisfactory results obtained if some of

(a) Complete common-base equivalent circuit and approximate high-frequency equivalent circuit.

(b) Common-emitter equivalent circuit and high-frequency approximation.

FIG. 24. Equivalent circuits and high-frequency approximations.

the resistive elements are dropped from the equivalent circuit. In Fig. 24 the symbols inclosed in parentheses are those sometimes applied to the elements of the equivalent circuit.[2] These circuits are only a few of the many that have been derived to describe the transistor and apply particularly to alloy and surface-barrier types. For some grown-junction transistors the base impedance is complex and for these types r_b' would be replaced by an impedance z_b'.

The equivalent circuits are approximations based on theoretical models. More exact equivalent circuits would involve transmission lines and would not be practical for engineering application.

23. High-frequency Parameters. The important parameters are f_α (alpha cutoff frequency), which should be high, r_b' (base spreading resistance), and C_c (collector-to-

[1] Figure 22*b* and *c* from Crow, R. P., Design of a High-fidelity 10 watt Transistor Audio Amplifier, IRE Convention Record, Part 7, *Audio and Broadcasting*, p. 142, 1956.

[2] Scarlett, R. H., "Some New High-frequency Equivalent Circuits for Junction Transistors," Stanford Electronics Laboratory Report 103, Mar. 20, 1956.

base capacitance), both of which should be low. An approximate expression for high-frequency gain for the common-emitter configuration for transistors having a constant r_b' is[1]

$$G_p \approx \frac{\alpha_0 f_\alpha}{25 f^2 C_c r_b'} \qquad 0.1 < \frac{\omega}{\omega_\alpha} < 2$$

and for grown junctions having a complex base impedance the high-frequency gain is

$$G_p \frac{0.2\omega_\alpha^{1/2}}{\omega^{3/2} C_c (r_b' r_e')^{1/2}} \qquad 0.1 < \frac{\omega}{\omega_\alpha} < 2$$

For both types the gain is dependent on the same parameters but the frequency behavior differs. For the types having a constant r_b', the high-frequency gain drops off 6 db per octave and for the grown-junction type having a complex base impedance the high-frequency gain decreases 4.5 db per octave.

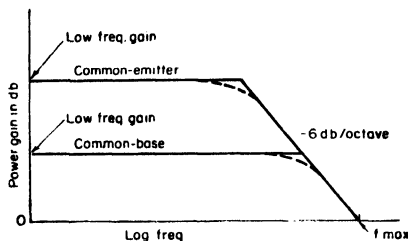



FIG. 25. Single-frequency gain vs. frequency.

24. High-frequency Gain. The low-frequency matched gains for the three basic transistors are, in order of gain, common-emitter, common-base, and common-collector. At high frequencies gains approach the same value and decrease at a 6 db per octave rate. The frequency at which gain is 1 (0 db) may be determined from the expression for gain. This is the maximum frequency at which the transistor can oscillate. The gain vs. frequency may be sketched by drawing a line from $f_{max} = \sqrt{\alpha_0 f_\alpha / 25 C_c r_b'}$ with a -6 db per octave slope to the point where the horizontal line representing the low-frequency gain is reached, as in Fig. 25.

25. High-frequency Impedances. As may be seen from the equivalent circuits, the input and output impedances are functions of frequency (see Fig. 26)[2] and bias conditions. For example, the emitter resistance r_e' is an inverse function of emitter current and the collector capacitance C_c varies as $V_c^{-1/2}$ for alloy or fused-junction transistors and as $V_c^{-1/3}$ for grown-junction units (V_c is collector-to-base voltage).

FIG. 26. Output capacitance and resistance vs. frequency for typical transistor in common-emitter configuration.

26. High-frequency Stability. All three transistor configurations have regions where they are potentially unstable, *i.e.*, passive terminations may be found to cause the circuit to oscillate. The common-collector and common-base configurations are potentially unstable over the whole high-frequency region with the greatest instability exhibited by the common-collector circuit. The common-base circuit has a region of potential instability from about $(1 - \alpha)\omega\alpha$ to a critical radian frequency

[1] Pritchard, R. L., High-frequency Power Gain of Junction Transistors, *Proc. IRE*, **43**, September, 1955.

[2] Chu, G. Y., Unilateralization of Junction Transistor Amplifiers at High Frequencies, *Proc. IRE*, August, 1953, p. 100.

of $r_e'\omega\alpha/2.4r_b'$, above which it is unconditionally stable.[1] Because of the stability problem unilateralization or neutralization circuits are frequently used in transistor r-f and i-f amplifier circuits. Some grown-junction units are perfectly stable at i-f frequencies yet have good high-frequency performance because of the combination of

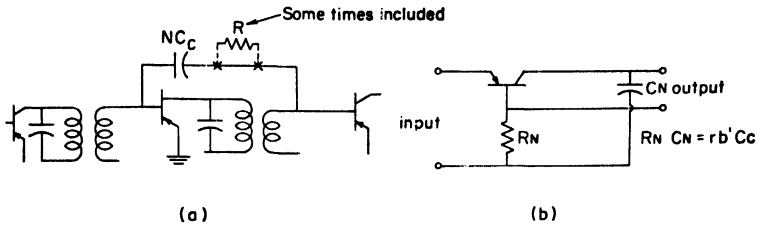



FIG. 27. Neutralizing circuits. (a) Common-emitter; (b) common-base.

a high base resistance resulting in a low critical frequency and a very low collector capacitance yielding a low $r_b'C_c$ product.

27. Neutralization. Figure 27 shows two of the more frequently used neutralizing circuits. Neutralization also eliminates the dependence of input and output imped-

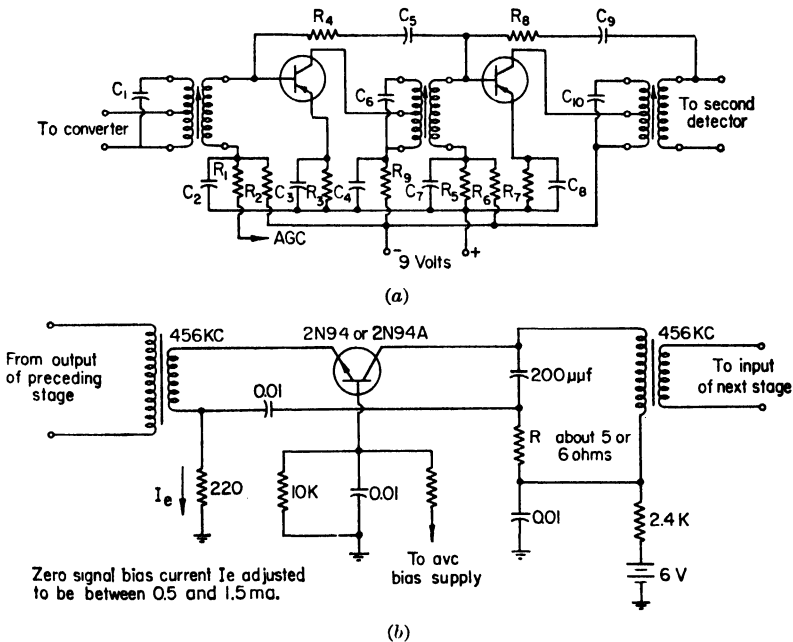



FIG. 28. Neutralized i-f amplifiers. (a) Neutralized as in Fig. 27a, RCA 2N139; (b) neutralized as in Fig. 27b, Sylvania 2N94.

ances on load and source impedance, respectively, thus simplifying alignment of tuned circuits. I-f amplifiers using neutralizing methods of Fig. 27 are shown in Fig. 28, which also shows the method of automatic gain control commonly used.

[1] Pritchard, R. L., Measurement Considerations in High-frequency Power Gain of Junction Transistors, *Proc. IRE*, **44**, August, 1956.

FIG. 29. Experimental

28. Gain Control. Gain may be controlled by reducing the emitter current with a signal obtained from the detector as the signal level increases. The reduction in emitter current causes an increase in emitter resistance and a reduction in gain. The decrease in gain is due to the decreased amplifying capabilities of the transistor and from a mismatch due to increased input impedance as the emitter current is reduced. Gain may also be controlled by a reduction in collector voltage but the emitter-current control appears more popular. The control signal is generally applied to the base, thus utilizing the d-c gain of the transistor in controlling the emitter current. Difficulties with automatic gain control include changes in band width with signal level and distortion at reduced emitter current, and because of this latter problem AGC is confined to low-level stages. In converter or mixer circuits optimum gain is obtained at a particular value of emitter current and gain control may be achieved by either increasing or decreasing the emitter current about the optimum value.

Single-tuned transformers having tightly coupled secondary windings are the most frequently used method of coupling in transistor i-f circuits. They provide a simple means of obtaining the desired impedance match between the output of one stage (about 50,000 ohms for a common-emitter stage) and the input impedance of the following stage (about 500 ohms for common-emitter). The secondary winding supplies the phase reversal for neutralizing as in Fig. 27*a*. For better selectivity double-tuned circuits, capacitively or inductively coupled, are sometimes used. Gains of 30 db are typical in i-f stages.

For wide-band amplifiers a double-tuned transformer having a capacitor in the secondary resonating with the leakage inductance at center frequency is a good method of coupling and may be designed for a maximally flat response. In common-emitter circuits the primary circuit is often tapped to have a reasonably small tuning capacitor (as in Fig. 27*a*); in common-base circuits the parallel output resistance is higher and desired output Q (about 35 for 456-kc broadcast receiver i.f.s) may be obtained without a tapped primary (as in Fig. 27*b*).

29. Typical Receiver Circuits. Figure 29 is a circuit diagram of an automobile receiver.[1] The r-f and oscillator circuits are permeability-tuned. The i-f stages utilize double-tuned capacitive-coupled interstages and are neutralized. Figure 30 shows a 20-Mc wide-band amplifier having a double-tuned transformer interstage.[2] The primary of the transformer is tuned with the output capacitance of the neutralized

[1] Freedman, Stanley, and Holmes, An Experimental Automobile Receiver Employing Transistors, *Proc. IRE*, June, 1955, p. 971.

[2] Scarlett, R. M., "Some New High Frequency Equivalent Circuits for Junction Transistors," Stanford Electronic Laboratory Report 103, Mar. 20, 1956.

automobile receiver.

common-base transistor and the 36-$\mu\mu$f capacitors tune the leakage inductance of the transformer to the center frequency. The transformers are designed to give a maximally flat band pass.

Circuit diagrams of some commercial receivers[1] are shown in Fig. 31. Circuits *a* and *b* of Fig. 31 use diodes as detectors and the source of AGC signal. In Fig. 31*a* the AGC signal is applied to the converter in a polarity so as to increase the emitter current with an increase in signal level. The AGC signal in Fig. 31*b* is applied to the base of the first i-f and to the combined oscillator and converter from the emitter of the first i-f stage with a polarity so as to decrease the emitter current with an increase in signal (note *n-p-n* transistors used in oscillator-converter and i-f stages). The circuit of Fig. 31*c* has a transistor for a detector, as does the circuit of Fig. 29, but has no

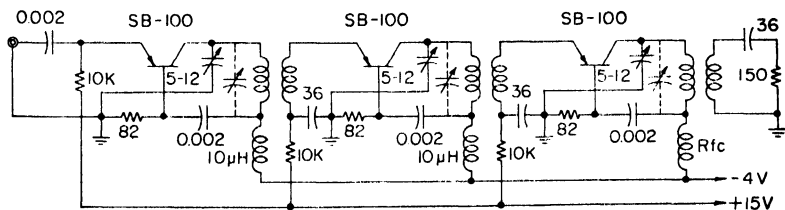



FIG. 30. Wide-band three-stage common-base amplifier.

AGC. The transistor detector has the advantage of giving gain as well as functioning as a detector and supplies power for AGC. The emitter-base diode is given a slight forward bias to bring the operating point near the knee of the emitter-base characteristic. In the circuit of Fig. 29, the AGC signal is taken from the emitter circuit of the transistor detector and applied to the emitters of the r-f and first i-f stages so as to decrease the emitter current with an increase in signal level.

In the circuits of Figs. 29 and 31, the oscillators all operate by inductive feedback from collector to base and in Fig. 31*b* and *c* one transistor performs the function of oscillator and converter. In the circuits of Fig. 31 the r-f signal and oscillator voltage are both applied to the base. In Fig. 29 the r-f signal is applied to the base and the oscillator signal is applied to the emitter. Gains of the order of 15 db are typical of converters. The oscillator coil in receivers is tapped to reduce the effective capacitance of the transistor appearing in shunt with the oscillator tuning capacitor, which

[1] Prensky, S. D., Current Trends in Transistor Radios, *Electronic Design*, July 15, 1956, p. 48.

FIG. 31. Commercial transistor broadcast receivers. (*a*) Raytheon, (*b*) RCA, (*c*) General Electric.

is placed across the primary, thus keeping the minimum capacitance obtainable with the tuning capacitor low and permitting the use of a small tuning capacitor. Other oscillator configurations are shown in Fig. 32. Oscillator circuits using vacuum-tube triodes have their transistor counterparts.

FIG. 32. Oscillator circuits.

Figure 33[1] shows the upper end of the frequency character-

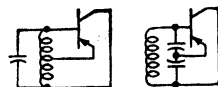
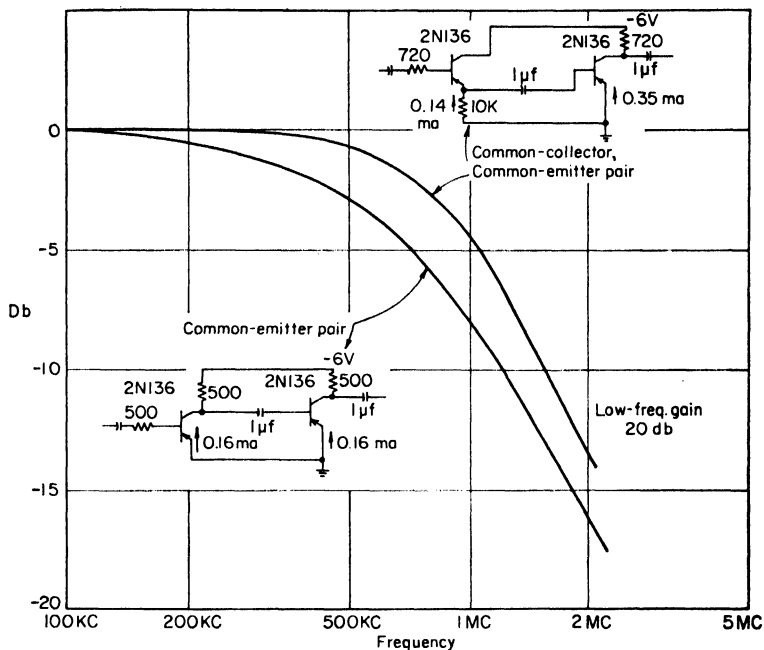



FIG. 33. High-frequency response of two video amplifier circuits.

istic of two video amplifier circuits, both having the same low-frequency gain of 20 db. The common-collector, common-emitter pair has a wider band width than a common-emitter pair. Figure 34 is a circuit[2] of a video amplifier using chokes to improve the band width. The amplifier

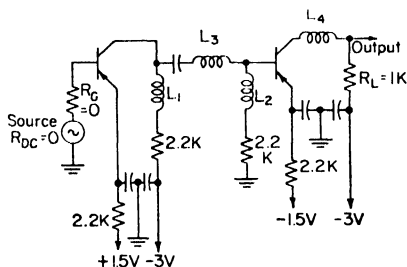



FIG. 34. Video amplifier, 9 Mc, 28 db, using surface-barrier transistors.

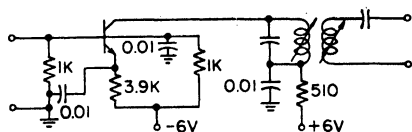



FIG. 35. Tetrode transistor amplifier; 7 Mc, 500 kc band width, gain > 18 db.

has a gain of 28 db, a band width of 9 Mc with surface-barrier transistors having an alpha cutoff of 45 Mc.

[1] Scarlett, *op. cit.*

[2] Angell, J. B., and F. P. Keiper, Circuit Applications of Surface-barrier Transistors, *Proc. IRE*, December, 1953, p. 1710.

For low radio frequencies (up to a few megacycles) a large number of grown and alloy-type junction transistors are available. Surface-barrier transistors are suitable up to about 20 or 30 Mc. Tetrode transistors are available in both silicon and germanium for amplification at frequencies somewhat higher than 100 Mc. Figure 35 shows the circuit diagram of a tetrode r-f circuit having a double-tuned transformer interstage.[1] Diffused-base transistors have gain beyond 200 Mc but are not presently available.

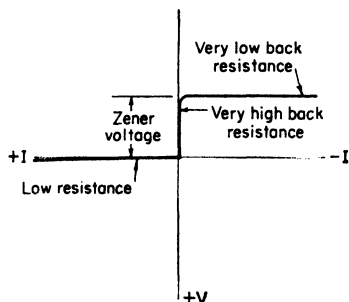



Fig. 36. V-I characteristic, silicon reference diode.

POWER SUPPLIES AND REGULATOR CIRCUITS

30. Advantages of Semiconductors. The greatest impact of semiconductor devices on vacuum-tube circuitry, probably, has been in the power-supply sections of electronics equipment. Some of the advantages over tube power supplies are reduction in size due to the semiconductor devices themselves, the absence of such bulky components as filament transformers, the minimization of heat generation of the rectifying units due to very low internal resistance of the semiconductor diodes (as low as 1 ohm or less); the increased ruggedness of the supply due to the inherently stable physical structure of the power transistor and power diode; and increased reliability of the supply due to these characteristics and the stable electrical characteristics of the power-type semiconductor units.

The most commonly used power diodes to date have been the fused-junction germanium or silicon units. A typical fused-junction silicon diode may have a peak inverse voltage rating of 500 volts and maximum average forward currents up to 1 amp and

Fig. 37. Shunt-type semiconductor constant-voltage regulators. (*a*) Low power; (*b*) single transistor; (*c*) output voltage control; (*d*) three transistors.

with internal resistances in the order of 1 ohm. The typical full-wave power supplies built with these units are of the 400-watt variety. For higher-wattage power supplies the individual units can be connected in series or parallel combinations (or large single units, now available, can be used). For very high wattage circuits, fan or liquid-cooled stacked units are used.

Before considering the subject of regulated power supplies, one must keep in mind that, because of the very low forward internal resistance of the semiconductor, one

[1] Schimpf, L. G., Tetrode Transistor Amplifiers, IRE-AIEE Conference on Transistor Circuits, February, 1954, pp. 18–19.

can build a brute-force-type supply with an internal impedance of a few ohms and this supply, in combination with a constant-voltage transformer, provides fairly good regulation features by itself. However, as the dimensions of the power transformer and chokes are usually prohibitive, the regulator-type circuit design is preferred.

There is a great similarity between the tube regulator circuit and the semiconductor regulator circuit. Where one would use a V-R tube or battery for a constant-voltage reference source in a tube regulator circuit, one would use a "zener" breakdown silicon diode or battery in a semiconductor circuit. Instead of a power pentode as the main regulating device, one would use a power transistor and instead of a high-gain tube for the control and feedback element, one would use a high-gain low-power transistor. The action of the "zener" voltage diode in providing a constant-voltage output over a wide variation of current through it results from the breakdown of a silicon-junction diode when biased in the back direction (see Fig. 36), thus providing the same effect as a V-R tube. The voltage breakdown can be controlled by the manufacturer to occur from a few volts up to a few hundred volts. The current range at each of these voltages is controlled by the maximum power dissipation of the

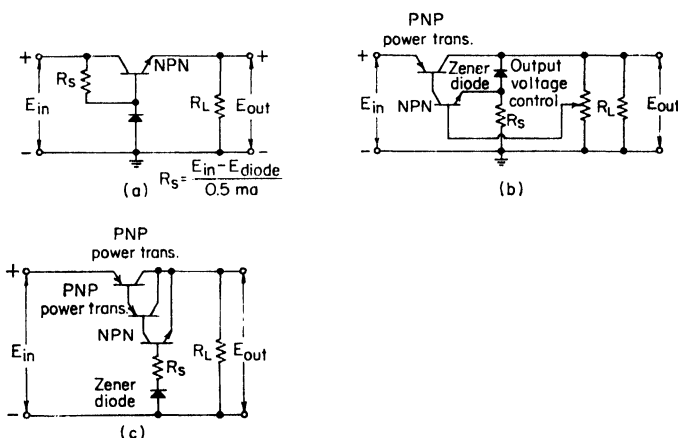



FIG. 38. Series-type regulators. (*a*) Low power; (*b*) output voltage control; (*c*) three transistors.

units, which is usually 100 mw. Figures 37 and 38 show some shunt and series regulator circuits using diodes and transistors. Figure 39 illustrates the use of semiconductor devices in a large power supply.[1]

TRANSISTOR SWITCHING CIRCUITS

The well-known and well-publicized characteristics of semiconductors make the transistor and diode of tremendous importance in the field of computers. Transistor computers can boast of very small power consumption, small size, and excellent reliability. It must be noted that the transistor cannot yet compete with the speed of vacuum tubes, but the difference in speed between the transistor and tube becomes less each year and in time the transistor will undoubtedly surpass the tube.

31. Binary-system Notation. Electronic computers perform mathematical calculations in a very basic, almost clumsy way. Computers, for example, do not use a decimal system with a base of 10 but use a base of 2.

The decimal number[2] system uses 10 different numbers or symbols which, in appro-

[1] Popp, F. G., "Transistor Regulated Power Supplies," Lincoln Laboratory Report 106, MIT, Jan. 19, 1956.

[2] From B. Dudley.

priate order or arrangement, can be made to represent any finite quantity. The binary number system, on the other hand, uses but two symbols, a 0 and a 1. Any number can be expressed in the binary system of notation, as well as in the decimal system.

FIG. 39. Regulated power supply for magnetic core memory.

In the decimal system, the position of the number indicates the power of 10 which is involved, and the size or magnitude of the digit indicates how many such numbers are to be considered. Thus, the number 751 really means

$$7 \times 10^2 + 5 \times 10^1 + 1 \times 10^0 = 700 + 50 + 1 = 751$$

In the binary system, the position of the symbol indicates the power of 2 involved,

and the symbol indicates whether a 0 or 1 is associated with that power of 2. Thus, when expressed in the binary notation system, 1110 means

$$1 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 = 8 + 4 + 2 + 0 = 14$$

The relation between binary and decimal numbers may be made more clear from the following table of equivalents:

Binary Number	Decimal Equivalent
0	0
1	1
10	2
11	3
100	4
101	5
110	6
111	7
1,000	8
1,001	9
1,010	10
1,011	11
1,100	12
1,101	13
1,110	14
1,111	15
10,000	16
etc.	etc.

32. Flip-flop Circuits. The basic building block of electronic computers is the flip flop. The flip flop is a circuit with two stable stages, corresponding to the two symbols used in the binary system. Figure 40 shows a standard transistor flip flop used in early transistor computers.

The cross-coupling resistors R_c and the collector load resistors must comply with the following condition[1]

$$\frac{R_c}{R_L} < \frac{2\alpha - 1}{1 - \alpha}$$

for stability. If this relationship is met, the collector voltage of transistor T_1 will be at one level while the collector voltage of transistor T_2 will be at another level. The collector voltage levels may be reversed by application of pulses to the transistor bases. A logical conclusion is that, by observing the original collector voltage levels, one can tell whether there were an odd or even number of input pulses to the transistor bases. This is a basic mathematical use of the flip flop. A means of counting input pulses will be discussed later, but a discussion of the operation of the flip flop will be presented first to provide a better background.

Referring to Fig. 40, assume that T_1 is "on" and T_2 is "off." The collector voltage of T_2 may be calculated using the equivalent circuit in Fig. 41. T_1 can be included in the above equivalent circuit simply by inserting its equivalent input resistance, which is in parallel with R_{b_1}. The input resistance of T_1 (at the base) is $R_e/(1 - \alpha)$. From the above equivalent network we can calculate V_{c_2} as

$$V_{c_2} = \frac{-V_{cc}\left[R_c + \dfrac{R_b \times R_e/(1-\alpha)}{R_b + R_e/(1-\alpha)}\right]}{R_c + R_L + \dfrac{R_b \times R_e(1-\alpha)}{R_b + R_e(1-\alpha)}} = \frac{-V_{cc}\left[R_c + \dfrac{R_bR_e}{R_b(1+\alpha) + R_e}\right]}{R_c + R_L + \dfrac{R_bR_e}{R_b(1-\alpha) + R_e}}$$

Similarly the collector voltage of T_1 can be found from the equivalent circuit of Fig. 42. The problem is solved more easily by separating the calculation into two parts. First the emitter current of T_1 is found and then using the relationship that $I_c = \alpha I_e$ the effect of collector current I_c on V_{c_1} can be found. In addition V_{c_1} is also part of a simple voltage divider as shown in Fig. 43. The collector current of T_1 is included in the voltage divider to

[1] The following derivations are based on the work of R. E. McMahon, "Transistor Flip-flop Design Techniques," Lincoln Laboratory, *Technical Memorandum* M24-Y2.

show its effect. That part of V_{c_1} due to the voltage divider is the first term on the right of the equation below and that part of V_{c_1} due to I_c is the second term on the right.

$$V_{c_1} = \frac{-V_{cc}(R_{c_1} + R_{b_2})}{R_{c_1} + R_{b_2} + R_{L_1}} + \frac{I_{c_1}R_{L_1}(R_{c_1} + R_{b_2})}{R_{c_1} + R_{b_2} + R_{L_1}}$$

The emitter current I_{e_1} and hence I_{c_1} may be calculated from the equivalent circuit in Fig. 44. The base circuitry can be reduced to a simplified circuit by use of Thévenin's theorem. The equivalent base voltage is

$$V_{beq} = \frac{-V_{cc}R_{b_1}}{R_{b_1} + R_{c_2} + R_{L_2}}$$

The equivalent base resistance is

$$R_{beq} = \frac{R_{b_1}(R_{c_2} + R_{L_2})}{R_{b_1} + R_{c_2} + R_{L_2}}$$

By combining the above a new equivalent circuit (Fig. 45) is obtained. The emitter current is found from

$$I_e = \frac{V_{beq}}{R_e + R_{beq}(1 - \alpha)}$$

Substituting the values for V_{beq} and R_{beq}, I_e becomes

$$I_e = \frac{\dfrac{-V_{cc}R_{b_1}}{R_{b_1} + R_{c_2} + R_{L_2}}}{R_e + \dfrac{R_{b_1}(R_{c_2} + R_{L_2})(1 - \alpha)}{R_{b_1} + R_{c_2} + R_{L_2}}}$$

Using these values,

$$V_{c_1} = \frac{-V_{cc}(R_{c_1} + R_{b_2})}{R_{c_1} + R_{b_2} + R_{L_1}} + \frac{I_{c_1}R_{L_1}(R_{c_1} + R_{b_2})}{R_{c_1} + R_{b_2} + R_{L_1}}$$

as given before.

The normal resistor values used in the circuit result in a voltage level for V_{c_1} of about -5 volts while V_{c_2} is about -10 volts or lower. Complete design equations for a flip flop of this type are found in the references at the end of this chapter. The -5-volt level can be designated as a "0" and the -10 volts can be defined as the "1" symbol.

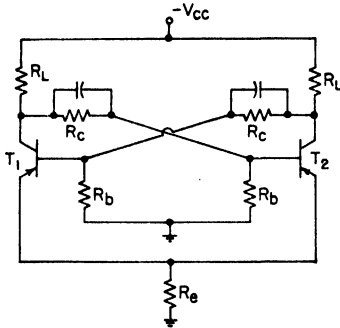



Fig. 40

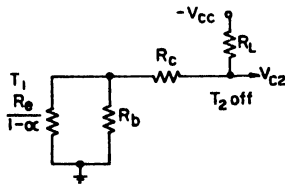



Fig. 41

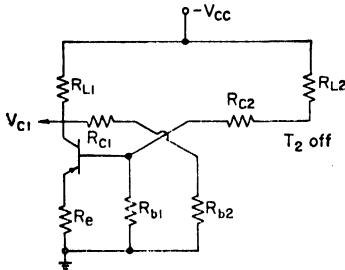



Fig. 42

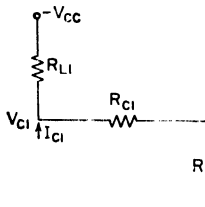



Fig. 43

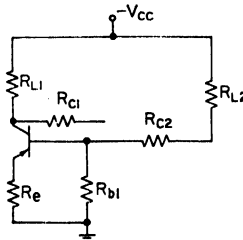
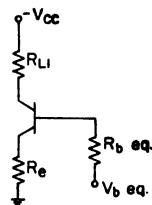



Fig. 44 Fig. 45

Figs. 40–45. Multivibrator development.

The above derivations indicate the general method of approach to transistor switching-circuit analysis. A derivation of the base voltages is also important since these voltages, relative to the emitter voltage, determine the stability of the flip flop. In the circuit of Fig. 40 the base voltage V_{b_1} of T_1 must be negative with respect to the emitter voltage V_e for T_1 to be "on" and the base voltage V_{b_2} of T_2 must be less negative than V_e for T_2 to be "off." In notation form, this indicates that

$$V_{b_1} < V_e$$
$$V_{b_2} > V_e$$

These relationships may be simplified to

$$V_{b_1} < V_{b_2}$$

If the voltage relationships V_{b_1} and V_{b_2} are derived and substituted in this equation the stability relationship will result, that is,

$$\frac{R_c}{R_L} < \frac{2\alpha - 1}{1 - \alpha}$$

If the resistor values are chosen in accordance with this equation then the flip flop will have two stable states. Since many resistance values hold for this equation, another relationship is needed to specify either R_c or R_L. Resistor R_L affects the collector voltage levels for both the "on" and "off" transistor, and so it is generally set. A simple relationship useful in determining R_L is

$$R_L = \frac{V_{cc}}{I_e} \times 0.8$$

I_e can be chosen to comply with transistor ratings and amount of driving current desired. The multiplying factor of 0.8 is based on advanced design relationships.

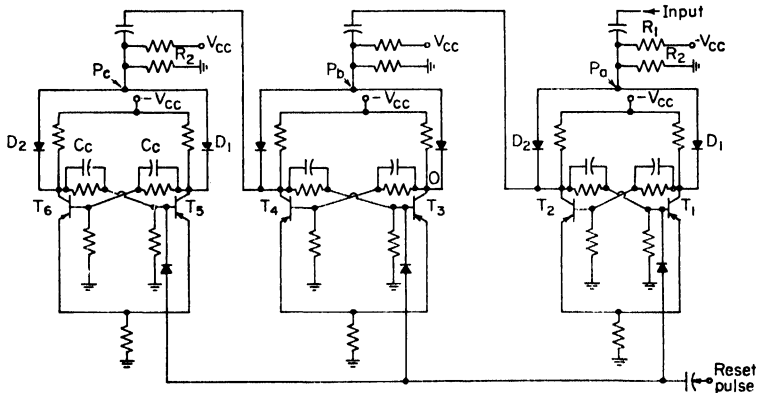



FIG. 46. Three-stage flip-flop counter circuit.

Once R_L is chosen, the stability equation may be solved for R_c. The minimum alpha expected for the transistors should be used. R_e affects the emitter current and may be calculated to give the emitter current selected.

The application of a positive pulse to the base of the "on" transistor (T_1) results in a reduction of emitter current and, hence, a reduction of collector current. The collector voltage level of T_1 drops because of the loss of collector current. This voltage drop is coupled to the base of T_2, causing an increase of T_2 emitter current. The emitter current of T_2 builds up, causing the voltage at the collector to rise. This

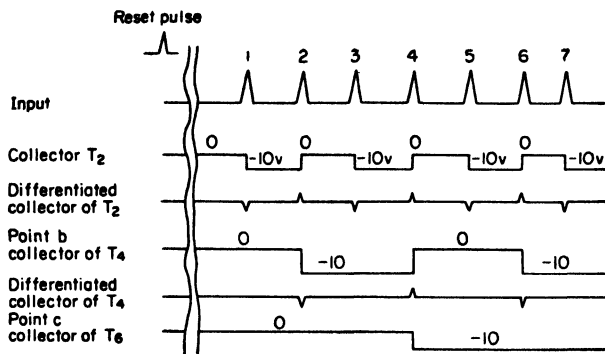



FIG. 47. Wave shapes for circuit of Fig. 46.

T_5	T_3	T_1	before first input pulse
0	0	0	before first input pulse
0	0	-10	first input pulse
0	-10	0	second input pulse
0	-10	-10	third input pulse
-10	0	0	fourth input pulse
-10	0	-10	fifth input pulse
-10	-10	0	sixth input pulse
-10	-10	-10	seventh input pulse
0	0	0	eighth input pulse

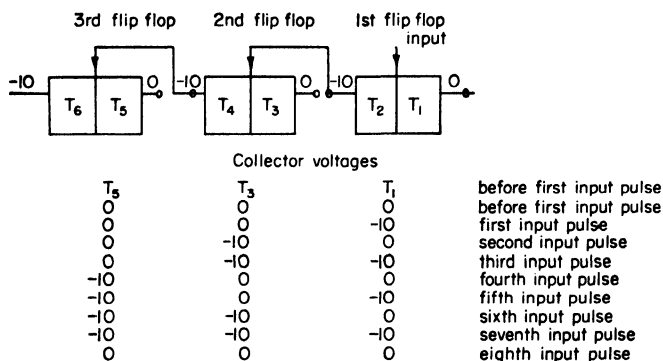



FIG. 48. Block diagram and coding of three-stage flip-flop counter.

(a)

A	B	C
0	0	0
0	-10v	0
-10v	0	0
-10v	-10v	-10v

(b)

A	B	C
0	0	0
0	-10v	0
-10v	0	-10
-10v	-10v	-10

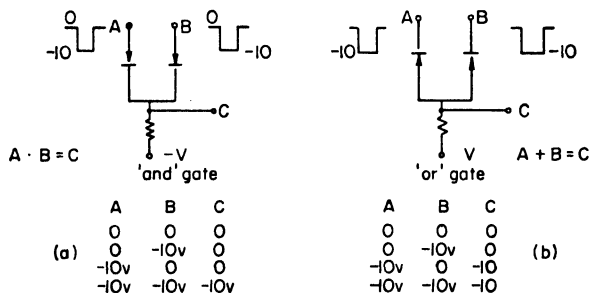



FIG. 49. Typical "and" and "or" gates.

voltage rise is coupled to the base of T_1, reinforcing the original pulse. After the regeneration is complete the state of the flip flop is changed so that T_2 is "on" and T_1 is "off." The application of a second pulse to the base of T_2 will cause the flip flop to change to the original state with T_1 "on" and T_2 "off."

A diode steering network can be added to the flip flop to steer a chain of pulses alternately to one base and then to the other base. Figure 46 shows a serial counter employing flip flops and diode steering networks. The voltage at the intersection of

the diodes (*Pa*) is set by R_1 and R_2 to equal the "off" transistor collector voltage level. A positive pulse applied to the input will pass through the diode connected to the "off" transistor collector and is coupled through C_c to the base of the "on" transistor, turning the "on" transistor "off." The next input pulse will pass through the opposite diode since it is now connected to the "off" transistor collector. The

Table 7. Desirable Characteristics for Computer-type Junction Transistors

Characteristics	Desirable conditions	Characteristic of germanium units	Characteristics of silicon units	Reason for desirable conditions	Typical triode junction transistor values	
					Germanium	Silicon
Alpha cutoff frequency $f_{\alpha co}$	High	High	High	Higher $f_{\alpha co}$ one of the factors contributing to higher frequency of operation (*i.e.*, faster read-out speed)	2–50 Mc	2–50 Mc
Max collector voltage breakdown $V_{c\,max}$	High	Medium	High	Higher $V_{c\,max}$ makes it easier to drive high-voltage indicators and sources, allows operation from very low V_c region where noise might be a problem	6–40 volts	30–100 volts
Base current amplification factor β at high values of emitter currents	High	High	Medium	Higher value of β at high currents makes for ease in core driving	15–60	5–30
Saturation resistance*	Low	Low	Medium	Low saturation resistance minimizes power dissipation in saturated-type circuitry	30–300 ohms	100–400 ohms
Extrinsic base resistance r_b'	Low	Low	Medium	The lower r_b', the lower the internal losses in emitter follower pulse amplifier, also aids in higher-frequency switching action	30–300 ohms	30–1,000 ohms
Emitter diode reverse breakdown voltage	High	High	Medium	Should be high enough to avoid clipping action† in large single switching applications	20–50 volts	5–20 volts
Collector capacitance C_c	Low	Low	Low	Low value of C_c aids in high-speed switching	7–20 $\mu\mu$f	3–12 $\mu\mu$f
Collector reverse leakage current I_{co}	Low	Low	Very low	Low values of I_{co} ensure maximum stability of operation	1–10 μa	<0.1 μa

* Resistance between the collector and emitter when both are biased in forward direction.
† Diode placed in series with emitter lead of silicon transistor will remedy this problem if encountered.

second pulse is coupled to the base of the "on" transistor and turns it "off." The output is coupled to the input of the second flip flop. The positive portion of the differentiated output causes the second flip flop to change states. The timing diagram of Fig. 47 indicates the sequence of events.

It is common to set the flip flops of a serial counter to a known state before the

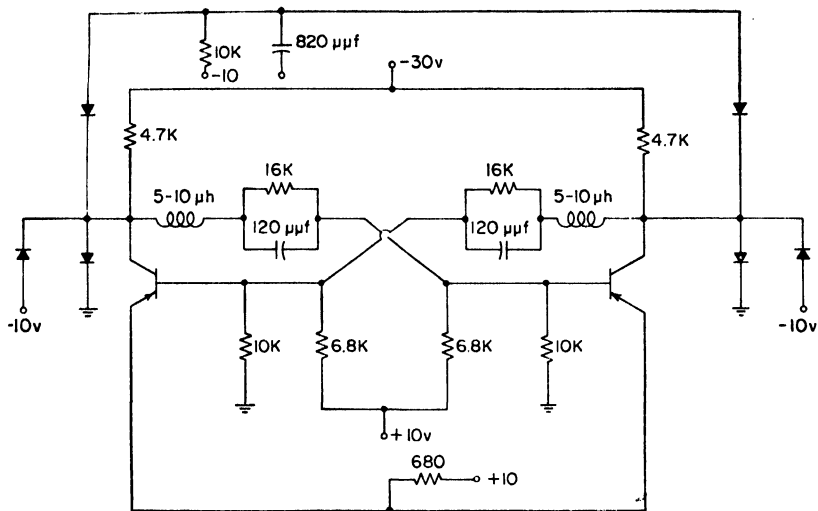



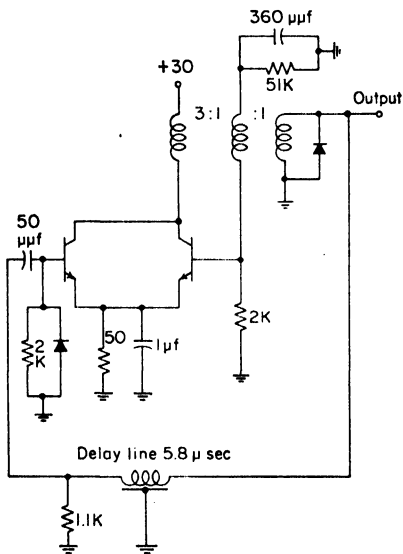
FIG. 50. Fast-flip-flop 5-Mc circuit.

application of the input pulses. In Fig. 47 the flip flops were assumed to be set so that T_1, T_3, and T_5 were on. The collector voltage levels are assumed to be 0 volts for the "on" transistor and −10 volts for the "off" transistor. A tabulation of the collector voltages corresponding to the input pulses is shown in Fig. 48. The tabulation may be verified from the timing diagram of Fig. 47.

If we allow the −10 volts to represent the "1" symbol and 0 volts to represent the "0" symbol, it can be seen that the tabulation above corresponds to the binary numbering system. In other words, after 3 input pulses the tabulation shows 0, −10, −10, or 011, which is the number 3 of the binary system. The three-stage counter will count up to 8 and then will repeat. A general formula for the number of stages necessary to count up to x is $x = 2^n$, where n is the number of stages.

The flip flop and serial counter are of basic importance in computers, but additional circuitry, such as shift registers, adder networks, blocking oscillators, and multivibrators is also important. In all these circuits transistors are applicable and are presently in use.

33. Other Switching Circuits. A brief description of transistor circuits other than the flip flops and serial counters is necessary to give a complete picture of switching circuits.

FIG. 51. Typical beat oscillator with self-pulsing controlled by delay line.

"And" and "or" gates are used in many ways, and a description of these gates will indicate their uses. Notice that the "and" gate (Fig. 49) will allow an output

(−10 volts) only if both *A* and *B* terminals are at −10 volts. In the "or" gate, an output occurs at terminal *C* if either *A* or *B* or both are at −10 volts. These gates can be used as control elements and are a necessary part of switching circuitry.

High-speed flip-flop and blocking-oscillator circuits are shown in Figs. 50 and 51.

Although the transistor switching circuit is basically an "on-off" device and well adapted to the binary system, other number systems may be used. For example, a counter made up of three flip-flop stages will count to 8 and could be used in systems

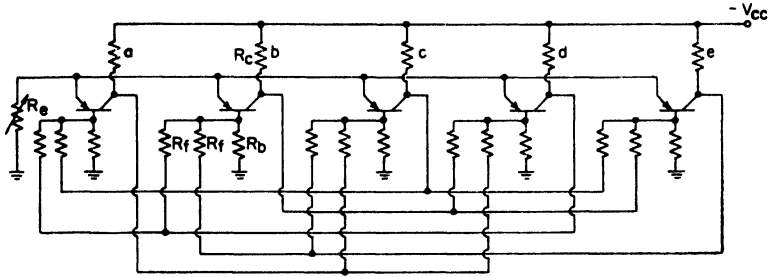



FIG. 52. Basic diagram ring-of-5 counter for junction transistor.

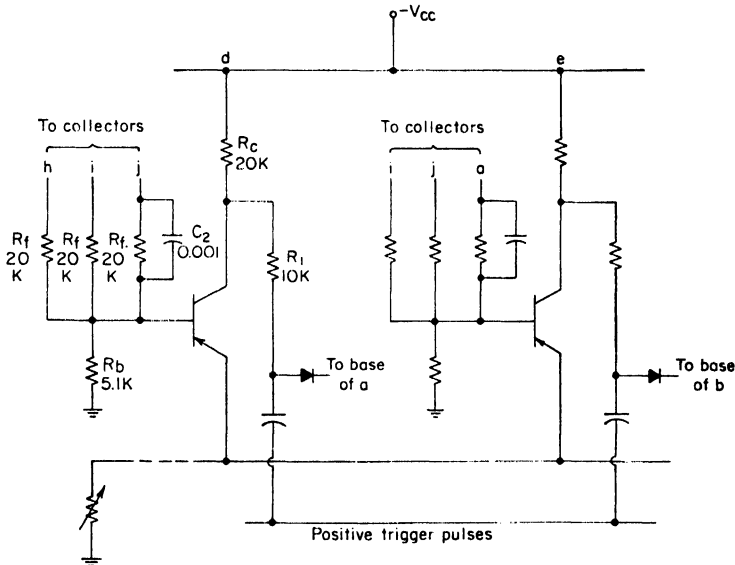



FIG. 53. Gating circuits for the ring-of-5 counter.

operating on base 8, or a four-stage flip-flop counter capable of counting to 16 may be restricted by additional circuitry to cycle at a count of 10, thus operating on the conventional decimal system. Ring counters provide further means of operating with number systems other than binary. Figure 52 shows the basic circuit of a ring of 5 using junction transistors.[1] This is not the conventional circuit commonly used in ring counters seen in tube literature where only one device in the ring is "on" at any time but is one in which, for a five-stage ring, two adjacent transistors are "on"

[1] Carlson, A. W., "A New Ring Counter for Junction Transistors and Vacuum Tubes," *Air Force Cambridge Research Center Report* AFCRC-TN 54-100.

and the rest are held "off" by the "on" transistors. Counters having one transistor (or tube) "on" and the rest "off" would require four resistors of the type labeled R_f and a five-stage ring would be the practical upper limit. By permitting other transistors in the ring to be "on" in a fixed and recognizable pattern of "on" and "off" stages, rings of 10 become practical. The count is indicated by the location of the adjacent "on" transistors in the ring with the location of these "on" stages being shifted by one stage at each successive count. The circuitry for transferring the position of the "on" stages in a 10-stage transistor ring counting positive pulses is shown in Fig. 53.[1] The letters in the diagram indicate the stages labeled a to j. In a 10-stage ring of this type four adjacent stages are "on" and the remaining are "off" at any time. Other types of ring counters may be made up of flip-flop stages and require two transistors per count capability.

Transistor circuitry at present has developed by applying vacuum-tube ideas and procedures and while this may have been a necessary step in the transition period it is not the most fruitful method. At present, more and more use is being made of characteristics peculiar to transistors, and considerable work is being done to change the basic circuitry to agree with transistor characteristics.

References

Adler, R. B., M. Boisvert, C. R. Hurting, and S. J. Mason: "A Large Signal Equivalent Circuit for Transistor Static Characteristics" in "Some Notes on Transistor Circuits," *M.I.T. Tech. Report* 25.

Angell, J. B., and R. P. Keiper, Jr.: Circuit Applications of Surface-barrier Transistors, Part III of The Surface-barrier Transistor, *Proc. IRE*, **41**, 1119, September, 1955.

Bright, R. L.: Junction Transistors Used as Switches, *Trans. AIEE*, Part I, Communication and Electronics, March, 1955, p. 111.

——— and A. P. Kruper: Transistor Choppers for Stable D-C Amplifiers, *Electronics*, April, 1955, p. 135.

Carlson, A. W.: "A New Ring Counter for Junction Transistor and Vacuum Tubes," AFCRC-TN-54-100, Air Force Cambridge Research Center, Bedford, Mass.

Chow, W. F., and A. P. Stern: Automatic Gain Control of Transistor Amplifiers, *Proc. IRE*, **43**, 1119, September, 1955.

Giacoletto, L. J.: Study of P-N-P Alloy Junction Transistor from DC through Medium Frequencies, *RCA Rev.*, **15**, 506, December, 1954.

Hilbourne, R. A., and D. D. Jones: Transistor Power Amplifiers, *Proc. Inst. Elec. Engrs.* (*London*), **102,** 763, November, 1955.

Keonjian, Edward: Temperature-compensated DC Transistor Amplifiers, *Proc. IRE*, **42**, 661, April, 1954.

Kruper, A. P.: Switching Transistors and as Substitutes for Mechanical Low-level Choppers, *Trans. AIEE*, Part I, Communications and Electronics, March, 1955, p. 141.

Loofbourrow, K. E.: Class B Operation of Transistors, I, Design Considerations, *Electronic Design*, July, 1955, p. 28; II, Basic Circuits, *Electronic Design*, August, 1955, p. 34.

McMahon, R. E.: "Transistor Flip-flop Design Techniques," Lincoln Laboratory, *Technical Memorandum* M24-Y2.

Mooers, Howard T.: "Design Procedures for Power Transistors," Minneapolis Honeywell Regulator Company, *Technical Report* TR77C, February, 1956.

Pritchard, R. L.: Electric Network Representation of Transistors—A Survey, *Trans. IRE*, **PGCT,** 12, March, 1956.

———: Measurement Considerations in High-frequency Power Gain of Junction Transistors, *Proc. IRE*, **44,** 1050, August, 1956.

———: High-frequency Power Gain of Junction Transistors, *Proc. IRE*, **43,** September, 1955.

Scarlett, R. M.: "Some New High-frequency Equivalent Circuits for Junction Transistors," *Tech. Report* 103, Mar. 20, 1956, Stanford Electronics Laboratories, Stanford University.

Schwartz, S.: Transistor Characteristics for Circuit Designers, *Electronics*, January, 1956, p. 161.

Shea, R. F.: "Principles of Transistor Circuits," Wiley, 1953.

Stern, A. P., C. A. Aldridge, and W. F. Chow: Internal Feedback and Neutralization of Transistor Amplifiers, *Proc. IRE*, **43,** 839, 1955.

Wheeler, A. J.: Thermistors Compensate Transistor Amplifiers, *Electronics*, January, 1957, p. 169.

[1] *Ibid.*

CHAPTER 11

LOUD-SPEAKERS AND ROOM ACOUSTICS

By Hugh S. Knowles[1]

In the design and operation of electroacoustic devices, consideration must be given both to the physical or "objective" properties of the sounds that are to be reproduced and to the psychophysiological or "subjective" processes involved in hearing.

1. Sound is an alteration in pressure, particle displacement, or particle velocity propagated in an elastic material or the superposition of such propagated alterations.

2. Sound is also the sensation produced through the ear by the alterations described above. In case of possible confusion the term "sound wave" may be used for concept 1, and the term "sound sensation" for concept 2.[2]

In the case of a sound wave in air the pressure is alternately above and below atmospheric.

The velocity of propagation, c, of a sound wave of small amplitude is

$$c = 33{,}060 + 61\theta \qquad \text{cm/sec} \tag{1}$$

where θ is the temperature in degrees centigrade. The wavelength λ is given by the relation $\lambda = c/f$, where f is the frequency in cycles per second. The density ρ of dry air at 20°C and at a pressure of 760 mm is 0.001205 g per cc.

The intensity of a plane or spherical "free" sound wave (no reflection) in the direction of propagation is

$$I = \frac{p^2}{\rho c} = 2.42 \times 10^{-9} p^2 \qquad \text{watt/sq cm} \tag{2}$$

where p is the effective sound pressure (dynes per sq cm).

The standard reference intensity is 10^{-16} watt per sq cm. The intensity level in decibels of a plane or spherical free sound wave in the direction of propagation is

$$L_I = 10 \log_{10} (2.42 \times 10^7 p^2) \tag{3}$$

In a plane or spherical free wave the intensity is proportional to the square of the pressure. In this case the pressure level in decibels of a sound wave is defined as

$$L_p = 20 \log_{10} (p/p_0) = 74 + 20 \log_{10} p \tag{4}$$

where p_0 is the standard reference pressure of 0.0002 dyne per sq cm. Two pressures are said to differ by x db if 20 times the logarithm to the base 10 of their ratio is x. As in the analogous electrical case involving a field strength or voltage ratio, this is valid only if the impedances are identical so that the energy is proportional in both instances to the square of the respective pressures. For this reason Eq. (2), in general, does not hold in more complicated fields. On the other hand the use of Eq. (4) is justified in more complicated sound fields by the fact that the hearing sensation depends primarily on the sound pressure.

1. Speech. The variation in conversational speech power with frequency is shown in Fig. 1. The ratio of ⅛ sec peak to averaged power in 15-sec intervals is roughly

[1] Knowles Electronics, Inc.

[2] American Standard Z24.1, Acoustical Terminology.

20 db. In overloaded amplifiers such as are frequently used in public-address systems, or in intentionally peak-clipped or compression systems the ratio may range from 3 to 10 db. This ratio is important in temperature-limited loudspeakers (see Tests).

The distribution of energy with frequency is brought out differently in Fig. 2.

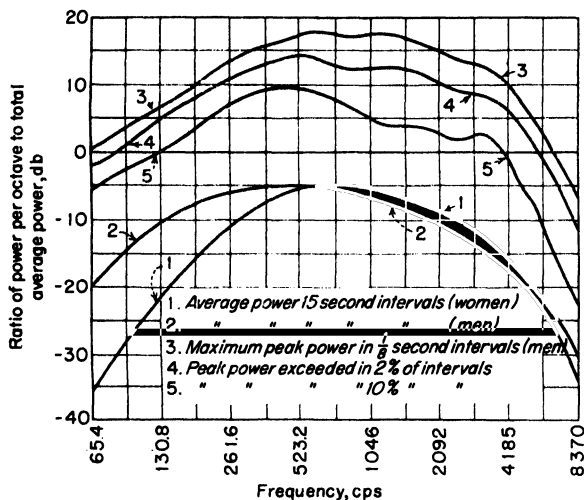



FIG. 1. Variation of conversational speech power with frequency. (*After Sivian and Fletcher.*)

Articulation curves which give a measure of the "recognizability" of speech are shown in Fig. 3. The percentage of called sounds correctly recognized is the per cent articulation. Tests of syllable, sound, vowel, individual sound, and other types of articulation are now widely used in the laboratory and to an increasing extent in the field to determine the suitability of a system for the transmission of speech. "Intelligibility" tests, in which the content of a simple sentence is to be understood, are also used. On the average 30 per cent syllable articulation corresponds to nearly 90 per cent "discrete sentence" intelligibility, indicating the relative ease of understanding connected speech. From Figs. 2 and 3 we note that reproducing only the frequencies above 400 cps halves the system power requirement and yet reduces the articulation by a negligible amount. In a power-limited system in which speech articulation is important, the transmission band is sometimes limited to from 600 or 800 to 4,000 cps, corresponding roughly to the 90 per cent articulation points at each end. This reduced band roughly quarters the power requirement.

FIG. 2. Speech-power variation with frequency. (*After Fletcher.*)

Articulation and naturalness are not to be confused. By successively raising the

cutoff of high-pass filters and lowering the cutoff frequencies of low-pass filters, each by a barely perceptible amount, Schäfer has shown that the required transmission band for natural speech reproduction includes some 32 to 36 minimum perceptible changes in band width.[1]

Gannett and Kenney have experimentally determined the successive reductions in the cutoff frequency of a low-pass filter required to obtain 75 per cent correct judgments by a trained jury when listening to speech and music.[2] On speech (average of male and female) the cutoff had to be lowered from the maximum available of 15 to 8 kc, then from 8 to 5 kc to produce a second reliably detectable step. On music the successive cutoffs were 11, 8, 6.5, and 5 kc.

FIG. 3. Variation of articulation with transmitted frequency range. (*After Fletcher.*)

Corresponding values for the successive increases in the cutoff frequency of a high-pass filter were calculated on less complete data. The cutoff had to be raised from 40, the lowest available, to 150 cps on speech for the first reliably detectable step. On music the first step was from 40 to 80 and the second from 80 to 150 cps.

FIG. 4. Variation with frequency of the power output of a 75-piece orchestra. (*After Fletcher.*)

From these data we conclude that, when the cutoff frequencies lie in the 40- to 150-cps and 5- to 15-kc ranges, trained listeners can detect twice as many band-width changes on music as on speech. Excellent speech can be transmitted over a 150- to 8,000-cps range and excellent music over an 80- to 11,000-cps range.

2. Music. The frequency distribution of the maximum and most probable peak powers for a 75-piece orchestra is shown in Fig. 4. The curves are based on average measurements of four selections which gave whole "spectrum" peak powers from 8 to 66 watts and average powers of 0.08 to 0.13 watt. Zero level corresponds to an average power of about 0.1 watt. As in the case of speech the

[1] Schäfer, E., The Audibility of Variations in Frequency Band in Speech Transmission, *Elek. Nach.-tech.*, **15** (8), 237, August, 1938.

[2] Gannett, D. K., and I. Kenney, The Discernibility of Changes in Program Band Width, *Bell System Tech. J.*, **23** (1), January, 1944.

average power over 15-sec intervals is about 1 per cent of the peak power in 18-sec intervals.

The power output of various musical instruments is shown in Table 1. The bass drum may radiate over a third of the peak power of a 75-piece orchestra. The large

Table 1. Peak Powers in Music

Instrument	Microphone position and assumption in converting to total sound power	Field pressure, dynes per sq cm		Total peak power, watts	Percentage of intervals	Frequency band containing max peaks, cps
		Average in 15-sec interval	Peak in ⅛-sec interval			
Bass drum, 36 × 15 in.	3 ft in front, on axis. Radiation confined to a cylinder having drum diameter	99.0	1,260.0	24.6	6.0	250– 500
Bass drum, 30 × 12 in.	Same as above	35.0	980.0	13.4	1.0	125– 250
Snare drum	4 ft in front, 90 deg off axis. Peak pressure increased 8.5 db for 1-ft distance. Radiation confined to hemisphere	14.6	365.0	11.9	2.5	250– 500
15-in. cymbals	3-ft distance. Peak pressure increased 7.2 db for 1 ft. Radiation confined to hemisphere	18.0	360.0	9.5	7.5	8,000–11,300
Triangle	3-ft distance. Conversion as for cymbals	2.3	25.8	0.05	1.0	5,600– 8,000
Bass viol	3-ft distance. Radiation confined to hemisphere	4.2	37.8	0.156	2.0	62– 250
Bass saxophone	3-ft distance. Radiation confined to hemisphere	4.1	58.2	0.288	25.0	250– 500
BB♭ tuba	3-ft distance. Conversion made from measurements with a complex sound source attached to a horn of similar size	5.4	43.2	0.206	17.0	250– 500
Trombone	3-ft distance. Conversion as for tuba	6.5	228.0	6.4	5.0	500– 700 2,000– 2,800
Trumpet	3-ft distance. Conversion as for tuba	8.6	54.2	0.314	18.0	250– 700
French horn	As for trumpet	3.8	27.0	0.053	6.0	250– 500
Clarinet	As for trumpet	3.3	26.4	0.050	5.5	250– 500
Flute	As for trumpet	1.6	25.6	0.055	1.0	700– 1,000 1,400– 2,000
Piccolo	As for trumpet	2.2	30.8	0.084	0.5	2,000– 2,800
Piano	10-ft distance. Room 29 × 29 × 13 ft. Reverberation time 1 sec 60–4,000 ~, average of 3 methods	2.6	23.4	0.267	16.0	250– 500
15-piece orchestra	6 ft from nearest instruments, in same room as piano. Average of 2 methods	7.9	126.0	9.0	1.5	250– 500 2,000– 2,800
75-piece orchestra	15 ft from nearest instrument in theater	4.6	129.0	66.5	1.0	250– 500 8,000–11,300
Pipe organ	Effective distance 15 ft. Radiation assumed uniform over ¼ sphere	20.0	90.0	12.6	36.0	20– 62.5

peaks in the 20- to 62.5 cps range of the organ are well known to recorders and electronic organ people who find it desirable to use l-f stops which are "rich in harmonic development" and therefore sound much louder without badly overloading the record, amplifier, and speaker. The 15-in. cymbals follow the drums and organ closely in peak power output with 9.5 watts. Their maximum peaks occur in the 8,000- to 11,300-cps range. Transmission systems having a "predistorted" frequency characteristic which includes a marked rise in h-f response in some part of the system (such as f-m and television transmitters) are frequently overloaded by this instrument. The same problem occurs in recordings recorded with a similar characteristic.

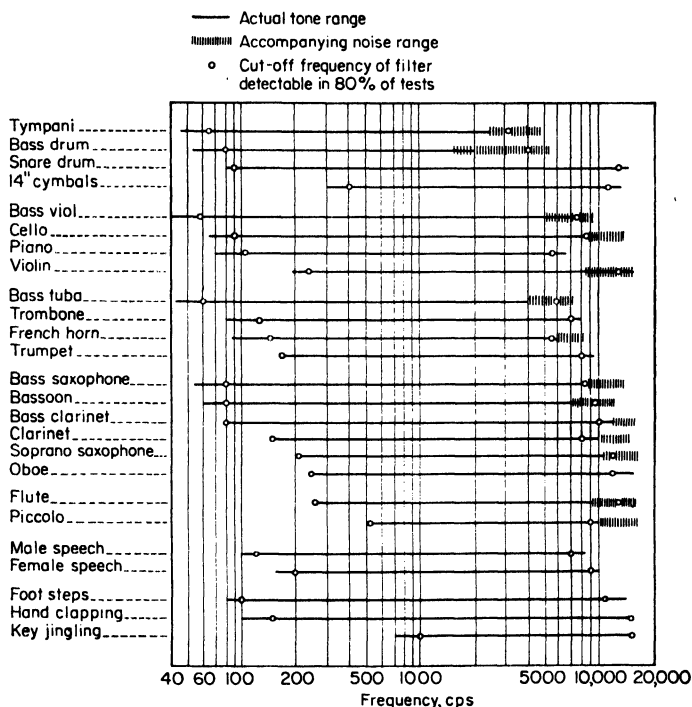



FIG. 5. Audible frequency ranges of musical instruments and sounds. (*After Snow.*)

The high output of the trombone in the 2,000- to 2,800-cps band near the frequency of maximum ear sensitivity gives the trombone (and other brass instruments) their piercing "bite." It has been found that the ear critically appraises the response of a system in this range and that surprisingly small changes can be detected. This suggests that the balance of the brasses in a studio pickup merits special attention.

The audible frequency ranges of many musical instruments are shown in Fig. 5. The vertical ruled portions indicate the frequency range in which noises accompanying the playing of the instrument occur. While the elimination of these frequencies permits the fact that the frequency range is restricted to be detected, it does not mean that the quality is judged to be best with the unrestricted range. In many cases the quality of the reproduced music from instruments which radiate extraneous noises (reed, bowing, key, and others) is improved by eliminating the noise range.

In restricting the transmitted frequency range of reproduced music, we have to be primarily concerned with the degradation in quality as judged by a good "sound jury" rather than with recognition of the selection played or the power distribution with

frequency or "spectral composition" of the music. The average results of a test of this kind, using a jury of 10 and an 18-piece orchestra, are shown in Fig. 6. Considering the many variables involved, the maximum and mean deviations from the curve were surprisingly small. It was the judgment of the observers that the quality improved rapidly as the lower range was extended down to 80 cps and the upper up to 8,000 cps.

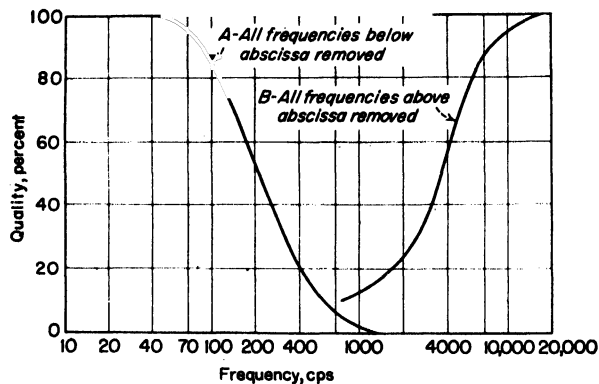



FIG. 6. Variation in quality of reproduced orchestral music with transmitted frequency range. (*After Snow.*)

It has been found experimentally that, if the transmitted frequency range is to be restricted, good balance between low and high frequencies may be obtained by so choosing the range that approximately equal degradation in quality occurs because of loss of low and high frequencies. For reasonable degradation the product of these two frequencies is roughly 640,000. The square root of this product or the geometric mean of these frequencies is therefore roughly 800 cps. A system transmitting more octaves below 800 cps than above usually sounds "heavy," "thick," or "drummy." Likewise a system transmitting more octaves above 800 cps than below will sound "thin" or "tinny." This assumes flat response

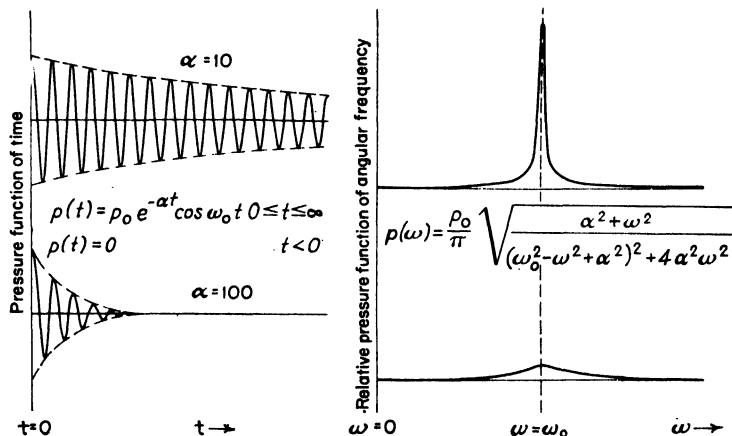



FIG. 7. Pressure variation with frequency for two isolated damped sinusoids with different rates of decay; initial amplitudes the same in the two sets of curves.

in the range and similar cutoff characteristics. A sharp cutoff at one end will increase the apparent output at that end because of the transient response which accompanies such a cutoff. A peak in either range will increase the steady-state and transient response in that region. This can be only partly balanced by added response in the other range.

In considering the problem of reproducing sounds in a complete system including

the effect of the room at the source of sound and at the source of the reproduced sound, it is desirable to know the energy distribution with frequency of a typical sound. The importance of this will be discussed under Room Acoustics. Since a common type of sound in music is a damped sinusoid, corresponding, for example, to the sound output of a plucked string instrument, the spectral analysis for two waves with different rates of decay is shown in Fig. 7. Any isolated wave train of this type contains energy which covers an infinite frequency interval. By analogy with the optical case the spectrum is said to be a continuous or band spectrum.

The highly damped wave contains appreciable energy at frequencies differing up to from 20 to 30 per cent from the frequency of a corresponding undamped wave. As the rate of decay is decreased, the wave train contains more energy, and an increasing amount of this is concentrated near the undamped frequency of the wave. In the limit when the rate of decay is zero and the wave has existed for an infinite length of time, *i.e.*, when we have a steady state, the band spectrum degenerates into a line spectrum with all the energy concentrated at the undamped frequency of the wave.

The fact that music and speech are not of a steady-state character but vary from instant to instant (and therefore have a continuous distribution with frequency of their energy) substantially aids their satisfactory transmission in a room (see Room Acoustics).

3. Noise. Noise is an "unpitched" sound composed of a large number of discontinuous, nonperiodic sounds. Therefore the energy in noise is distributed in a continuous manner with frequency. A click, for example, closely approaches the hypothetical pulse which lasts for an infinitesimal length of time and the energy of which is continuously and uniformly distributed with frequency. A noise may have one or more broad peaks in its spectrum, but a sharp peak indicates a nearly periodic disturbance which will give the noise a definite pitch.

Table 2. Noise Levels

Pressure, dynes per sq cm	Noise level, db above reference threshold	Type of noise
630	130	Pain threshold
250	122	Airplane—1,600 rpm, 18 ft
45	109	Boiler factory
25	102	Subway train passing station
13	96	Elevated train—15 ft
4.0	86	Heavy traffic—15 ft
2.0	80	Average truck—15 ft
1.3	75	Average factory location
6.3×10^{-1}	70	Average automobile—15 ft
3.2×10^{-1}	64	Department store
1.1×10^{-1}	55	Average office
2.8×10^{-2}	43	Quiet office
6.3×10^{-3}	30	Very quiet residence
1.4×10^{-3}	17	Gentle whisper—5 ft
4.5×10^{-4}	7	Threshold (for street noise)

The properties of noises are of some importance because (1) the proper reproduction of intended noises may enhance the dramatic value of the reproduction, (2) the ambient noise levels in studios or halls and in rooms frequently limit the dynamic range at the "pickup" and "playback" points, and (3) they influence the response of the ear by producing masking or artificial deafening.

The reproduction of most noises requires the transmission of substantially the entire audible frequency range. For this reason noises are frequently used as test material in high-quality systems. The frequency ranges of footsteps, hand clapping, and key jingling are shown in Fig. 5. These indicate that it is particularly important that all the upper audible frequencies be transmitted.

The intensity level of various representative noises is listed in Table 2. In urban locations, particularly in large buildings, the ambient noise level in moderately quiet rooms is of the order of 45 to 60 db. This noise level is high enough so that even in specially treated broadcasting studios it frequently limits the dynamic range of the transmitter.

Even in relatively quiet residential sections the ambient noise level in a typical listening room is high enough so that it, too, places a lower limit on the intensity of the sound required to override the noise.

4. Hearing. All the data contained under Speech, Music, and Noise which were obtained by the use of a sound jury or listener involve the sensation produced in the listener by the designated source of sound. All tests of this type depend to some extent on the techniques employed and, of course, on the observers. All similar tests are of principal value when the jury is composed of a large selected sample with known hearing characteristics.

One of the most important properties of sound is its loudness. In the case of typical sounds and especially those in the mid-frequency range the intensity (power) must be increased by a factor of 10 to double the loudness. In high-power systems

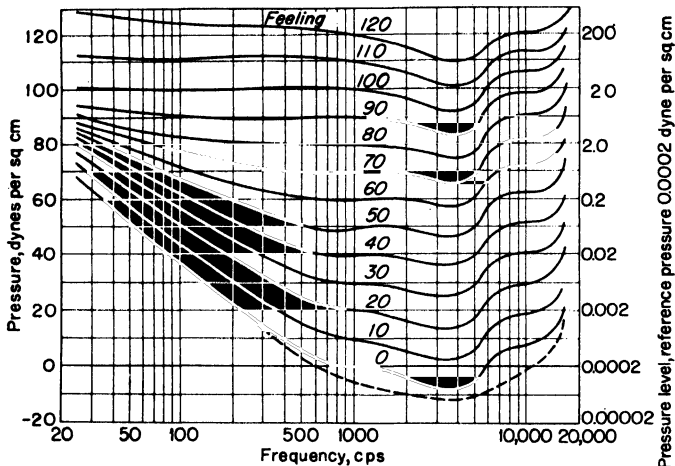



FIG. 8. Loudness level curves showing variation in sound intensity with frequency required to produce a sound judged to be as loud as the 1,000-cps reference sound intensity given on the curves. (*After Fletcher and Munson.*) Solid curves obtained with listener facing sound source. Dashed curve indicates threshold (corresponding to solid curve 0) but for sound of random incidence. (*After Sivian.*) For more recent data on a larger sample of ears see D. W. Robinson and R. S. Dadson, A Redetermination of the Equal Loudness Relations for Pure Tones, *Brit. J. Applied Phys.*, **7**, 166–181, 1956.

the required loudness should therefore be critically evaluated. This has been found to vary with both the frequency and intensity of the sound. To a rough approximation it has been found that in the middle frequency range equal percentage increases in intensity produce equal increases in loudness. The *loudness* is the magnitude of the hearing sensation and is assumed proportional to the number of nerve impulses reaching the brain per second.

The *loudness level* contours for a sample of 200 ears are shown in Fig. 8. These curves were obtained by alternately listening to a sound of arbitrary frequency and intensity and comparing it with a 1,000-cps tone the intensity of which was adjusted until the two were judged to be equally loud. At 1,000 cps therefore the loudness level of the sound corresponds to the intensity level because this is the reference test frequency. The intensity is that which exists in an undisturbed sound field before the listener is immersed in it. The observer faces the source and listens to the sound binaurally. By plotting the differences in minimum audible field intensities for sound of normal and random incidence found by Sivian, we obtain the dashed curve in Fig. 8. This indicates that the other contours for sound of random incidence would also be more regular.

Reports have been made by Beasley on a sample of 16,000 ears. Some of the results are shown in Fig. 9. The curves show the percentage of the sample tested

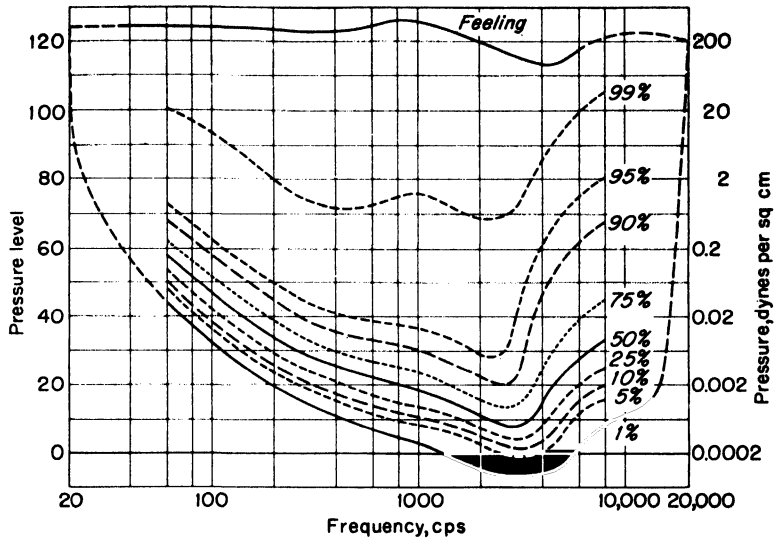



FIG. 9. Threshold of hearing curves for large population sample. Percentage figures indicate percentage of sample tested having a hearing threshold lower than the corresponding curve. (*After Beasley.*)

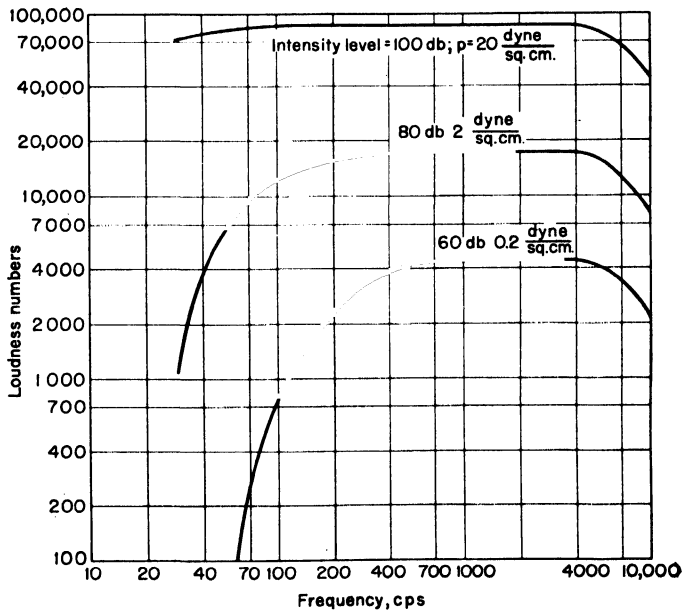



FIG. 10. Loudness variation with frequency for three tones of the indicated intensity showing reduced loudness of low-intensity low-frequency sounds.

which had lower thresholds of hearing than the indicated value. For example, the solid curve marked 50 per cent indicates that 50 per cent of the ears tested had thresholds of hearing lower than that indicated by this curve. From these data we see that the Fletcher and Munson threshold curves are for ears in the best 1 per cent of the 16,000-ear sample, and that hearing deficiencies are prevalent enough to justify their consideration in equipment design.

The loudness or apparent response or transmission characteristic of a system emitting a plane free sound wave of three constant intensity levels is shown in Fig. 10. A sound of constant 100-db intensity level seems almost equally loud from 30 to 6,000 cps. A sound of moderate intensity (constant 60-db intensity level) is inaudible below 60 cps and increases in loudness rapidly up to 400 cps. In the presence of noise, masking would substantially reduce the loudness at low intensities. The l-f characteristic varies much more rapidly with intensity level than the high and for this reason compensated volume controls are designed to have their maximum effect at low frequencies. It should be noted that the intensity-level compensated characteristic is a purely arbitrary thing and, although it is considered superior by some, it does not correspond to our normal experience. In practice, when we go some distance from the source, the low frequencies become inaudible. When an orchestra plays at low intensity, we get the same effect.

The effect of noise on hearing is to produce artificial deafness or "masking." The nature of the apparent deafness which results depends on the energy distribution of the noise. In normal listening environments the maximum noise *intensity* occurs at low frequencies. In fairly common cases the pressure decreases by a factor of 2 as the frequency increases by the same factor (slope = 6 db per octave). The effect of moderate noise levels is to decrease articulation. If the noise level is moderate this may be largely compensated by raising the intensity level of the sound.

LOUD-SPEAKERS

A loud-speaker is a device which is actuated by electrical signal energy and radiates acoustical energy into a room or open air. The shorter term *speaker* is used when no confusion with a person addressing a microphone results.

The selection and installation of a speaker as well as its design should be guided by the problem of coupling an electrical signal source as efficiently as possible to an acoustical load. This involves the determination of the acoustical load or radiation impedance and selection of a diaphragm, motor, and means for coupling the loaded loud-speaker to an electrical signal source. The performance of the speaker is intimately dependent upon the nature of its acoustic load and should not be considered apart from it. The nature of the radiating system, and therefore the acoustic load impedance it sees, is primarily determined by space, acoustical environment, and cost factors.

5. Radiation Impedance. When a vibrating diaphragm is placed in contact with air, its impedance to motion is altered. The added impedance seen by the surfaces which emit useful sound energy may be called the *radiation impedance.* By analogy with antenna systems the resistive part is called the *radiation resistance.* The radiation reactance or reactive part is usually positive, and the corresponding apparent mass may be called the *radiation mass.* The radiation impedance seen by a diaphragm depends on its size, shape, the frequency, the acoustical environment, and the medium into which it radiates.

6. Single Piston. The average radiation impedance per unit area seen by a flat circular piston vibrating in a thin, rigid, nonabsorbent, infinite plane or baffle in air is shown in Fig. 11. When the length of the radiated sound wave λ exceeds the circumference of the piston, $2\pi R$, the radiation resistance is nearly proportional to the square of the frequency. This relation between the wavelength and piston circumference corresponds to the frequency being less than 4,310 divided by the diameter in inches. In this frequency range the piston *velocity* should vary inversely with frequency to radiate constant power since this is equal to the product of the square

of the rms velocity and the radiative resistance [see Eq. (5)]. This variation in velocity with frequency is usually obtained by placing the fundamental resonant frequency of the diaphragm and motor near the lowest frequency to be transmitted so the system has mass reactance or is "mass-controlled" in this frequency range. When the frequency exceeds twice this value, the resistance is very nearly 41.4 mechanical ohms per square centimeter and the diaphragm (real or virtual) is efficiently coupled to the air (see Sec. 11).

When the length of the radiated wave exceeds the circumference of the piston, the air increases the apparent mass of each side of the diaphragm by approximately the

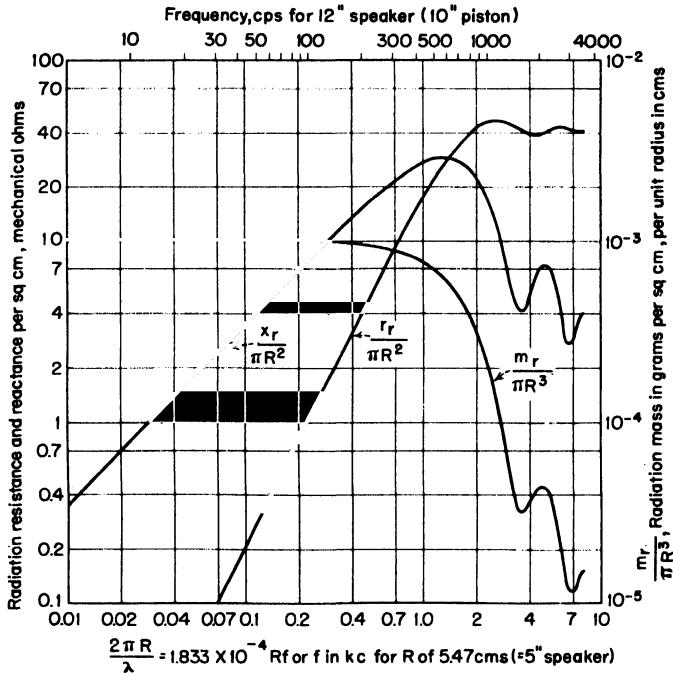



FIG. 11. Radiation resistance, reactance, and mass per square centimeter of a flat, rigid piston vibrating in an infinite, rigid nonabsorbing baffle. Piston radiates into a solid angle, $\Omega = 2\pi$ steradians (hemisphere).

mass of air contained in a cylinder whose base is the piston and whose height is 0.85 times the piston radius. At high frequencies the radiation mass ("accession to inertia") and the mass reactance decrease and approach zero for infinite frequency.

7. Mutual Radiation Impedance. When a sound wave radiated from one surface of a diaphragm has access to another surface of the same diaphragm or to a surface of another diaphragm, there is said to be coupling between the surfaces. Consideration of this mutual radiation impedance is simplified by fixing attention on what occurs at each diaphragm. The motion of the diaphragm is opposed by the ("self-") radiation impedance. It is also opposed or aided by the force exerted on it by the waves generated by any other diaphragms which are coupled to it. The (complex) ratio of the force due to all other diaphragms to the velocity of the diaphragm itself is the mechanical impedance seen by the diaphragm due to the other diaphragms. This we will call the total mutual radiation impedance.

The total radiation impedance seen by a diaphragm is the sum of the self- and the

"mutual" radiation impedances. The acoustic power P_a radiated by a diaphragm is

$$P_a = (r_S + r_M)v^2 \times 10^{-7} \quad \text{watt} \tag{5}$$

where r_S = real part of self-radiation impedance (total)
r_M = real part of mutual-radiation impedance (total)
v = rms diaphragm velocity, cm per sec

Note that the velocity appears as current would in the corresponding electrical equation. The diaphragm displacement is $v/2\pi f$, where f is the frequency. The total displacement of various diaphragms required to radiate 1 watt is shown in Fig. 12. These curves clearly show the need for large diaphragms if appreciable l-f power is to be radiated.

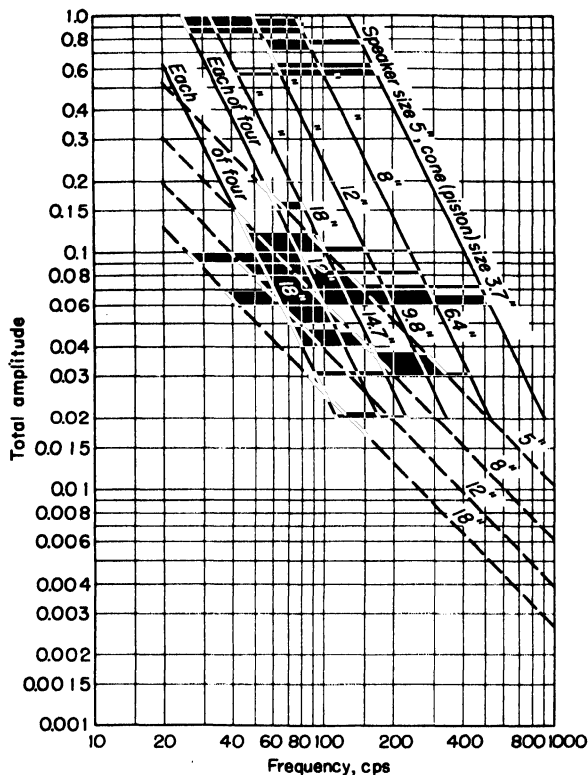



Fig. 12. Total displacement required of diaphragm of indicated size to radiate 1 watt. Solid curves for pistons radiating into a hemisphere at low frequencies where the radiation resistance is proportional to the square of the frequency ($2\pi R/\lambda$ less than about 1.4 in Fig. 11). Dashed curves for constant radiation resistance of 41.5 mechanical ohms per square centimeter (exponential horn value well above horn cutoff frequency).

By knowing the self- and mutual radiation impedances of diaphragms mounted in a single infinite baffle, we can determine the impedance seen when other baffles are added. In Fig. 13 assume four pistons (1, 2, 3, 4) mounted in the X plane. Because of symmetry there is no net sound flux through the plane Y since for every positive vertical (z) component from pistons 3, 4 there is a negative component downward from pistons 1, 2. We may therefore introduce the rigid, thin, nonabsorbent plane or baffle Y without altering the impedance seen by any of the pistons. With Y in place we may remove pistons 3, 4, and pistons 1, 2 will continue to see the same impedance. The sound wave reflected by the plane corresponds exactly to the wave which would come from diaphragms 3, 4 and therefore the plane is said to have created "primary images" (by analogy with the optical case) of diaphragms 1, 2 which 1, 2 cannot distinguish from the real diaphragms 3, 4. Similarly the plane Z may be introduced and pistons 2, 3, 4 removed, leaving 1 looking into its original impedance. In all cases pistons of equal size, vibrating in phase and with the same amplitude in infinite, rigid, nonabsorbing baffles, are assumed. The relations hold approximately when the baffles are a wavelength or more long. Finite

impedance of a baffle may be treated by assuming reduced amplitude of the image to account for absorption and a change in phase to account for the reactive part of the impedance. The principle is readily extended to multiple sources of arbitrary size, phase, and displacement such as occur in vented enclosures, labyrinths, and the like.

Several piston combinations are shown in Fig. 14. All pistons marked with the same letter "see" the same radiation impedance. The ratio of the radiation resistance and reactance seen in each case to that seen by a single piston A is shown in Figs. 15 and 16. The actual impedance is therefore obtained by multiplying the ordinates of Fig. 15 or 16 by the corresponding ordinate of Fig. 11. The letters on the curves correspond to those on the pistons in Fig. 14.

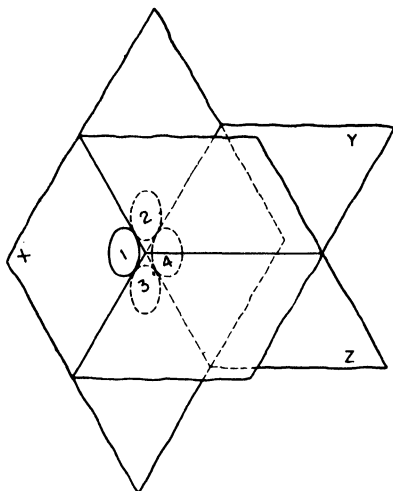



FIG. 13. Primary images 2, 3, and 4 of piston 1 introduced by planes Y and Z.

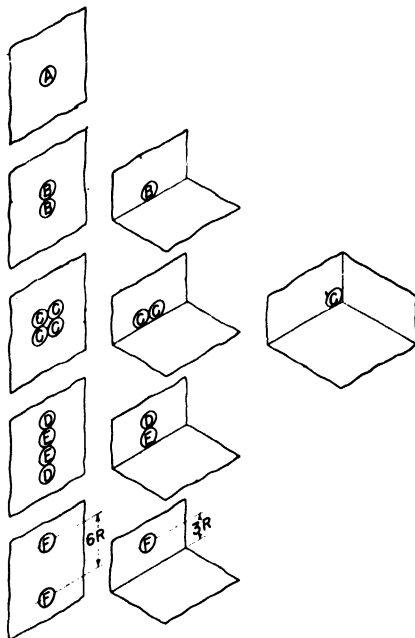



FIG. 14. Effect of adding pistons and reflecting planes on radiation impedance. All pistons marked with the same letter see the same radiation impedance.

Values of the ordinate less than 1 indicate the piston sees less resistance or reactance than it would if alone in a single infinite plane. This occurs when the time delay in the wave from one diaphragm and the frequency are such as to give out-of-phase pressure components at the other.

Figure 15 shows that the radiation resistance is increased by the largest factor and over the greatest frequency range when each diaphragm is as close to all others as possible. At low frequencies the group then behaves as a single large piston. Figures 15 and 16 show that the radiation resistance increases much more rapidly than the reactance as speakers (or their images) are added. The exact improvement in efficiency which results depends on the impedance seen looking back into the speaker diaphragm but a typical speaker efficiency is increased by a factor of nearly 2 (3 db) for case B and 3.2 (about 5 db) for case C. This indicates qualitatively the improvement gained at low frequencies by operating a speaker at the intersection of the floor and wall and in a corner, respectively.

For a given diaphragm amplitude one speaker in locations B and C will radiate two and four times as much l-f energy, respectively, as one in location A. The radiated

power for constant amplitude is therefore proportional to the square of the number of actual diaphragms. The radiation resistance at high frequencies is not improved by the use of additional speakers. A group of speakers therefore has better low but no better and possibly less h-f response than a single one, and they therefore sound as though they had relatively less high response. Except for cost reasons multiple speakers are usually preferred to a single speaker with the same size motor because (1) the small diaphragms are lighter per unit area than a large one of adequate rigidity, giving better efficiency and high response; (2) the angle of individual speakers may be adjusted a moderate amount to give a good h-f directional pattern without injuring the l-f response; (3) reliability is improved, since failure of a single unit usually does not seriously affect the performance of the group; and (4) the temperature rise of each voice coil may be reduced.

FIG. 15. Ratio of radiation resistance seen by a piston in the presence of one or more others (real or images) vibrating with equal amplitude and phase to radiation resistance it would see alone radiating into a hemisphere. Designating letter of curve is same as that of corresponding piston in Fig. 14. (*After Klapman.*)

MULTIPLE LOUD-SPEAKERS

Some of the advantages of multiple direct-radiator speakers where these all cover the same frequency range are discussed in Sec. 7, Mutual Radiation Impedance, above. It is also shown, however, that at intermediate and high frequencies, added diaphragms may, and usually do, reduce the mutual radiation resistance. This results in dips in the response curve of the combination of speakers, which may be serious. Multiple speakers are also sometimes used to control directivity as in "line" sources in which as many as 8 or 10 are mounted in a vertical

FIG. 16. Ratio of radiation reactance (or mass) seen by a piston in the presence of one or more others (real or images) vibrating with equal amplitude and phase to reactance (or mass) it would see alone radiating into a hemisphere. Designating letter of curve is same as that of corresponding piston in Fig. 14. (*After Klapman.*)

line to reduce the vertical angle of radiation but to preserve a broad horizontal angle. Multiple-speaker systems in which the speakers cover complementary frequency ranges also have certain advantages and are widely used. The more important advantages

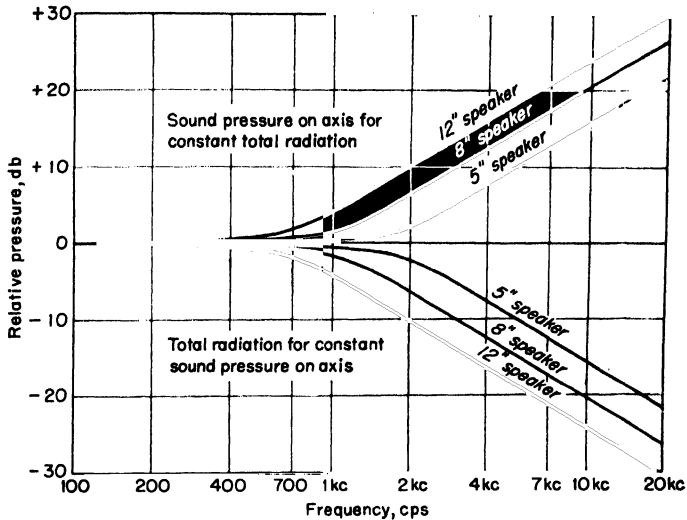



FIG. 17. Effect of concentration of radiation on piston axis at high frequencies for the case of pistons radiating constant *total* power (top curves) and the case of constant pressure response on the axis (lower curves). Because of cone flexing the concentration is less marked in actual diaphragms where the equivalent piston diameter at high frequencies approaches about six-tenths the actual cone diameter.

are (1) improved frequency response, since each type of unit covers a moderate range; (2) higher system efficiency, for the same reason; (3) improved directivity characteristic, since the diaphragm (or horn mouth) for the highest frequency range may be made relatively small (see Figs. 17, 18, and 19); (4) improved transient response, since many of the artifices used to obtain extended frequency ranges in single units make the transient response worse, particularly at high frequencies; (5) reduced intermodulation, since large amplitudes are confined to speakers reproducing low frequencies; and (6) reduced frequency modulation which occurs when a single diaphragm moves with large amplitude at low frequencies with respect to the listener while simultaneously radiating high frequencies, thereby altering the frequency of the high frequency because of the Doppler effect.

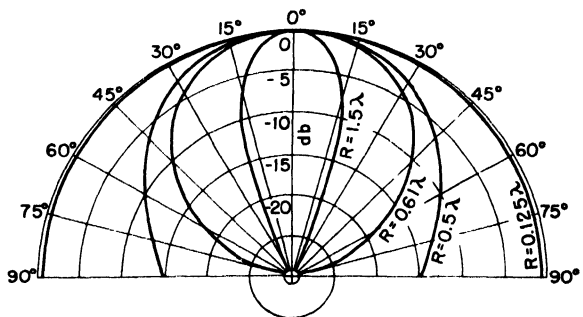



FIG. 18. Variation in relative response with angle up to the angle for which the first minimum occurs. The response on the axis has been arbitrarily adjusted to the same reference level in all curves.

8. Piston Directivity. With rising frequency the radiation from a rigid piston becomes increasingly concentrated on the axis, as shown in Figs. 17 and 18. Figure 17 shows that, if the flat axial pressure response curve so often sought after is obtained, the total radiation and therefore the efficiency are actually falling rapidly at high frequencies. Conversely, if a speaker is to have constant efficiency its axial pressure response must rise appreciably at high frequencies. The variation in relative response with angle up to the angle for which the first minimum occurs is shown in Fig. 18. The response on the axis has been arbitrarily adjusted to the same reference level in all curves. At high frequencies the effective area of an actual cone is reduced by flexing, so that the directivity of actual cones is somewhat less than that shown for the piston.

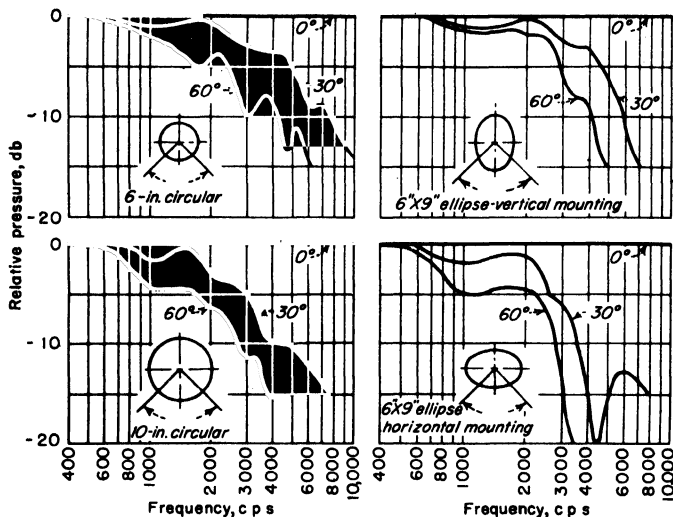



FIG. 19. Experimentally determined directional characteristics for one elliptical and two circular diaphragms. The elliptical diaphragm has the broadest characteristic in the *horizontal* plane when its major (long) axis is *vertical.*

Typical directional curves for 6- and 10-in. (designating size) speakers are shown in Fig. 19. The axial response is assumed equalized to give flat response and the relative response for other angles is shown. Typical directional curves for a 6- by 9-in. (designating size) elliptical speaker are also shown. These show that the directional response of this elliptical speaker in the plane of the *minor* or short axis is roughly comparable to that of a circular speaker with a diameter equal to the minor axis. The directional response in the plane of the major or long axis is worse than that of a 10-in. circular speaker up to about 6,000 cps. Above this frequency it is better. Contrary to popular belief the elliptical speaker should be mounted with its major axis *vertical* to get broadest distribution in the *horizontal* plane. This corresponds to the theoretical predictions of Stenzel. This same effect is present in rectangular mouth-shaped horns where the middle and middle h-f response is *most directional* in the plane of the *broadest* horn mouth dimension unless partitions, or separate cells, are used. Even then the effect is present when the wavelength is comparable to the smaller dimension of the entire mouth.

9. Horns. A horn is a tapered acoustical transmission line used to couple the impedance it sees, looking back into the diaphragm, as effectively as possible to the load it sees looking out of its mouth (see Motors). The small end of a horn is called its *throat,* the large end its *mouth.* If its mouth has an infinite flange or baffle, the

radiation impedance it sees is approximately the impedance given in Fig. 11. If there is no flange, the radiation resistance is half and the reactance approximately seven-tenths this value at low frequencies. At high frequencies the flange does not alter the impedance.

It has been found that many of the horns commonly used in loud-speakers belong to the hyperbolic exponential or "catenoidal" horn family, in which the cross-sectional area, A, is related to the axial distance by

$$A = A_t \left[\cosh\left(\frac{x}{x_0}\right) + T \sinh\left(\frac{x}{x_0}\right)\right]^2 \tag{6}$$

where A_t is the throat area, x_0 is a constant fixing the axial scale of length, T is a constant determining a member of the general family, and the cosh and sinh are the hyperbolic exponential cosine and sine functions, respectively. The longitudinal sections of these horns for various values of T are shown in Fig. 20 for straight-axis circular horns.

Fig. 20. Longitudinal section of straight-axis circular cross-sectional horns of hyperbolic exponential or catenoidal family.

When T is infinite, a conical horn is obtained. For T equal to unity the bracketed expression reduces to e^{2x/x_0}, yielding the equation for the familiar exponential horn. At T equal to zero the shape is that of the hyperbolic cosine, with zero slope at the origin.

The performance of a horn depends principally on the throat impedance and its dependence on frequency. While waves reflected from the mouth introduce fluctuations into the throat impedance, it has been found that the average impedance is closely that of a horn with no reflected wave. The throat impedance in mechanical ohms of catenoidal horns with rigid, nonabsorbent walls and negligible reflected waves is given by

$$Z_t = R_t + jx_t = A_t\rho c \frac{[1 - (f_c/f)^2]^{1/2} + j(Tf_c/f)}{1 - (1 - T^2)(f_c/f)^2} \tag{7}$$

where ρ, c, and T have been defined, f is a frequency, and f_c is the cutoff frequency, given by $f_c = c/2\pi x_0$. Thus the reference axial length x_0 is of fundamental importance in determining the behavior of the impedance. In Fig. 21 is shown the behavior of R_t for various values of T. The product ρc has been replaced by 41.4, its value in cgs units at room temperature.

Fig. 21. Frequency dependence of throat resistance for horns of various values of T. Below f_c the resistance is zero.

For frequencies below f_c the resistance R_t becomes zero, and the horn cannot lose energy to the medium. In practice there is a slight amount of radiation below this, so that the transition is not too abrupt. In many designs the cutoff of the system comprising amplifier, horn-driver unit, and horn may be judged aurally to be 10 or 20 per cent above the theoretical horn value because of impedance effects arising from other elements of the system (see Motors).

The behavior of the horn throat impedance is conveniently expressed by an equivalent electrical circuit of the mechanical elements, as in Fig. 22. The constant mass, m_t, depends inversely on T and is shunted

by the frequency-dependent resistance, R_t. These two elements account for the high-pass characteristic of horns. The driver unit is coupled to the horn by the sound chamber, which has a volume, V_c, terminated by the horn-throat area, A_t, and by the diaphragm area, A_d. This acoustic element may be represented by S_c, the shunt stiffness of the enclosed air plus an ideal transformer of turns ratio A_d/A_t. Referring again to Fig. 22 the diaphragm mass and stiffness, m_d and s_d, are driven by a source of force with internal impedance R_0 mechanical ohms. It is this "generator resistance" to which the horn must be properly coupled to attain a speaker of high efficiency.

The reactive elements in the equivalent circuit can be chosen on the basis of filter theory, but the response so obtained may not be the type desired. In general, the ratio A_d/A_t may be chosen so as to produce approximate equality of source and load resistances in the "pass" or transmission band. In this case the resonant frequencies

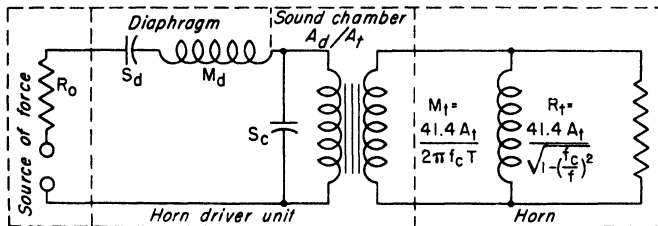



FIG. 22. Equivalent electrical circuit of mechanical elements of horn speaker. The generator impedance R_0 increases with the efficiency of the speaker.

of the diaphragm alone, and of the stiffness s_c with the mass $(A_d/A_t)^2m_t$, are placed so as to emphasize the desired frequency regions.

An important horn design parameter is the cutoff frequency, which must be properly related to the resonant frequency of the diaphragm to produce the desired l-f response. Once a motor and diaphragm have been selected, the horn throat area is usually fixed by the resistance-matching consideration mentioned, and T is chosen from Fig. 21. Thus the desired horn contour may be plotted from Eq. (7) and continued until the mouth is sufficiently large to minimize reflection. This occurs at a mouth diameter greater than about $\lambda/3$ at the low cutoff frequency. An equivalent statement is that the product of the cutoff frequency and the mouth diameter (in inches) is greater than 4,500. This product can be considerably reduced if the speaker is operated under matched resistance conditions with the mechanical source resistance R_0 equal to the mechanical resistance of the loaded diaphragm loaded by the horn for then a 4:1 change in load resistance results in but a 2-db reduction in output. This can be attained, over a broad frequency band, only with a very efficient motor. Thus the irregularities due to wave reflections at both ends of the horn do not show up strongly in the radiated energy.

The dimensions of the sound chamber are fixed by considerations of distortion and h-f cutoff. If the motion of the diaphragm is comparable to the clearances, then the stiffness, s_c, will vary throughout the cycle and lead to intermodulation distortion. However, if the clearance is increased, the value of s_c is lowered, hence its reciprocal, the capacitance, raised, thus shunting the horn load excessively at high frequencies. To calculate the amplitude of diaphragm motion there may be used the expression for the radiated power

$$P_a = 10^{-7}(\omega y_d)^2 A_d^2 \frac{R_t}{A_t} \tag{8}$$

evaluated about an octave above the cutoff frequency. Here P_a is the acoustic power in watts, and y_d the diaphragm amplitude in centimeters. Note that frequency enters into both ω and R_t.

At high frequencies such that the wavelength is comparable to the diaphragm diameter, transverse waves in the sound chamber produce cancellation effects which sharply reduce the radiated power. This condition can be ameliorated by subdivision of the throat into properly placed annular slots, as in Fig. 23, or by an annular diaphragm, as in Fig. 24. In these structures the sound chamber stiffness, s_c, may be considered a lumped constant equal to $1.37 \times 10^6 A_d/V_c$ dynes per cm over most of the frequency range, where A_d and V_c are defined above. The capacitance is the reciprocal of this value.

Fig. 23. Cross section of annular V-shaped diaphragm and sound chamber used to reduce destructive interference in sound chamber at high frequencies.

The cross-sectional area of exponential horns varies exponentially with length. Thus a horn with a 70-cycle cutoff doubles in throat area every 11 in. along the axis, and the area of a 100-cycle horn increases 20 per cent every 2 in.

With a horn the directionality of the speaker may be controlled almost independently of the response, because the directional properties are fixed mainly by the mouth size and shape. Thus it is possible to obtain in horn speakers more control over response, efficiency, and directivity than in direct-radiator types.

Fig. 24. Cross section of a diaphragm and sound chamber in which annular slots are used to reduce destructive sound-chamber interference at high frequencies.

DIAPHRAGMS

10. Principle of Operation. The diaphragm is the part of the speaker which couples the radiation impedance to the speaker motor impedance. In the usual hornless or direct-radiator speaker, the force exerted by the motor is localized, and this must be transmitted to the acoustic load which is spread over a large area. To do this effectively and to add as little as possible to the impedance, the diaphragm is made as rigid and light as possible. Conical shaped diaphragms are commonly used because they are rigid. The term "cone" is frequently used as a synonym for diaphragm in direct-radiator speakers.

The usual conical diaphragm may be thought of as a continuous mechanical transmission line radiating acoustic energy from each element of area. There exist both radial waves which travel from the driven region to the edge and are reflected, and circumferential waves which travel around the diaphragm. Both occur in various combinations depending on the mode of vibration and the symmetry of the driving force, diaphragm, and supporting members.

The lowest frequency mode and the simplest ane is the one in which the effective radial wavelength of the diaphragm, including the edge termination, is $\lambda/4$. (This must not be confused with a $\lambda/4$ in air at the same frequency.) At this frequency, which ranges from 700 cps in large to 1,400 cps or more in small conical diaphragms, no circumferential wave is present, and all parts of the cone move in phase. The

displacement is a maximum at the apex and a minimum at the flexible annulus which supports the outer edge and terminates the transmission line. The impedance of this termination plays an important part in the diaphragm behavior, especially at frequencies near the fundamental resonance of the diaphragm and motor and in the 1,000- to 2,000-cps range. At the l-f resonant frequency it may be considered a lumped stiffness whereas at the h-f end its distributed mass and stiffness are important.

At frequencies below the lowest mode of the cone itself all parts of the cone move in phase, and the cone behaves approximately as a piston unless the annulus stiffness increases rapidly with displacement, in which case the cone may flex at even very low frequencies. The annulus is frequently made this way deliberately in inexpensive speakers to produce l-f distortion and substantially increase their loudness by radiating most of the energy at harmonic frequencies. Unfortunately, intermodulation of low and high frequencies then also occurs, which makes the high end sound rough or garbled when a strong low note is reproduced.

11. Size. It has been found experimentally that at low frequencies the effective area of the cone is its projected or base area. This is approximately the "cone" size where this is defined for a circular cone as "the diameter to the nearest ¼ in. of the minimum circle determined by the tangency of the cone and a plane touching its base."[1] This is not to be confused with the *designating* size of a loud-speaker which is commonly used in describing a speaker.

> The designating size of a loud-speaker employing a circular radiator shall be twice the maximum radial dimension, measured to the nearest eighth-inch, of the front of the speaker except that the designating size shall not exceed the maximum diameter of the unsupported portion of the vibrating system by more than 25 per cent.[2]

This definition is intended to limit the amount of functionally useless cone housing included in the designating size. Representative cone sizes for various speaker-designating sizes are shown in Fig. 12.

FIG. 25. System efficiency of four speakers using identical moving-coil motors but different cone sizes (calculated). Designating size of speaker is given.

In direct radiator speakers and at low frequencies the radiation resistance is proportional to the fourth power of the radius (square of the area) and the reactance to the cube of the radius. The resistance-reactance ratio or power factor of the radiation impedance is therefore proportional to the piston radius. For constant radiated power, at a constant low frequency, the piston displacement varies inversely with area. With fixed amplitude the radiated power is proportional to the square of the area at a given frequency, or equal power may be radiated at half the frequency if the area is increased by a factor of 4. The upper limit to diaphragm size is set by the increased mass *per unit area* required to get a sufficiently rigid structure. The nature of the acoustic load (horn, enclosure, cabinet, etc.), space limitations, cost, and the motor employed also control size.

It is customary to increase the size of the motor as the diaphragm size is increased, since the mechanical impedance looking back into the motor (voice coil, say) should go up as the impedance looking into the driving point of the cone rises to maintain

[1] EIA, definition M5-111.
[2] EIA, definition M5-110.

good energy transfer. If a large cone is put on a small motor, the displacement and distortion for a given acoustic output drop and lower frequencies may be reproduced with the same distortion, but the efficiency in the mid-range may actually drop. These effects are illustrated in Fig. 25, in which the calculated system efficiency of four speakers using different size cones but the same motor is shown. Minimum cone masses, found to be satisfactory experimentally, and average mechanical resistance and resonant frequencies were assumed. The motor is an intermediate size normally employed on 8-in. speakers but frequently used on all four diaphragm sizes. Speaker efficiency, even at low frequencies, is therefore not limited by cone size. The cone size must be large, however, if appreciable power is to be radiated with reasonable cone excursions at low frequencies.

12. Shape. The most efficient shape at low frequencies is circular. This is also the most satisfactory structurally. Theoretical and experimental investigations have shown that an ellipse with a major to minor axis ratio of two and a two-to-one rectangle have an average of 5 and 7 per cent lower radiation resistance in the useful l-f range than a circle of the same area. The loss is progressively greater as the shape departs still further from circular. In spite of the visual appeal of elliptical and other diaphragm shapes, their fabrication and structural disadvantages have prevented their general adoption.

At high frequencies all pistons have the same radiation resistance per unit area, but most cones cannot be considered pistons, both because they are not flat and because their radial length exceeds $\pi/4$ and flexing is therefore important.

The shape of the cross section or profile of the diaphragm depends on the application and response desired. Straight side cones without corrugations are usually employed when high total radiation in the 2,000- to 5,000-cps range is required and when reproduction above 6,000 to 7,000 cycles may actually be undesirable. This is frequently the requirement of outdoor public-address and phonograph systems where noise and distortion are otherwise objectionable. Curved cones improve the response above 6,000 to 7,000 cps by providing a diaphragm impedance, viewed from the voice coil, which has a more uniformly high negative reactance and therefore absorbs more power from the high positive reactance (due to the voice-coil mass) seen looking back into the voice coil. This improvement is obtained at the expense of 2,000- to 5,000-cps response and with a weaker cone structure, with the result that straight cones are more commonly used.

13. Material. Hard, impregnated or filled, and pressed or calendered papers are used when loudness efficiency and apparent h-f response are important. Radiation resistance provides very little dissipation in direct radiator cones; hence, by using a paper having low internal flexural losses, the conical transmission line is made to have strong resonances. Many direct radiator speakers now use material of this type. The transient response of diaphragms of this type is necessarily poor since noncenter moving modes of the cone are inappreciably damped by the motor. Soft, loosely packed, or felted blotterlike cones are used when some loss in h-f response can be tolerated and a smoother response curve with reduced transient distortion is required. The loudness efficiency of high-loss cones of this type is several decibels lower than that of low-loss cones.

Felt, leather, rubber, and similar materials have been used as the annulus to terminate the conical transmission line in a low-stiffness high-resistance material. Their effect is to add considerable dissipation to the cone at the termination, resulting in reduced reflection of the flexural wave. The effect is similar to that obtained in soft cone materials, where, however, the dissipation is distributed along the line. These materials are now used infrequently because they increase the cost, lower the loudness efficiency, and necessitate larger air-gap clearances than are used when the annulus is made an integral part of the diaphragm. A more common method is to coat the paper annulus heavily with a plastic having high dissipation.

14. Breakup Subharmonics. The term *cone breakup* is sometimes applied to the flexing or wave-transmission process in a cone. Since there is nothing discontinuous

in the process to suggest the word "break" and to avoid ambiguity, it is suggested that this term be applied only in the other sense in which it is used, to name the process which results in the generation of subharmonics.

If the apex of a cone is driven with an adequate sinusoidal force at certain critical frequencies, the radiated wave contains not only the fundamental and integral multiples (harmonics) of it but also a frequency corresponding to half (and rarely to a quarter) that of the fundamental and integral multiples of this subharmonic. While distortion in the motor may contribute to this, some unpublished research has shown that the cone is the important source.

FIG. 26. Variation in radiated distortion products with voice-coil current showing abrupt start of half frequency (subharmonic) and odd multiples of it. Subscript indicates factor by which fundamental frequency is multiplied to obtain frequency of indicated distortion product.

The half frequency appears very suddenly at a critical input as shown in Fig. 26. To simplify the graph, the fundamental and usual harmonics, which would include even multiples of the subharmonics, are not included. From the total (rms) harmonic curve H_Σ, which includes these unplotted terms, we see that negligible rise in total distortion occurs when the subharmonic begins. The ear, however, reports a large increase because the pitch sense of the output has dropped an octave and the distortion has a high annoyance or objectionability factor. This type of distortion is not so important as is frequently supposed, however, because (1) it occurs only in limited frequency regions; (2) it does not occur below a moderate, critical level; (3) the time required to start it is large unless the force is large; (4) the energy distribution of speech and music is such that the probability of its production is small. Because of the statistical improbability of its frequent occurrence, it would be uneconomical to design most systems to avoid completely this occasional distortion.

MOTORS

A loud-speaker motor converts electrical into mechanical energy and couples the electrical signal source as efficiently as possible to the mechanical impedance seen looking into the diaphragm which it drives.

15. Force Factor. The mechanical circuit of a speaker motor experiences a force when a current is applied to the electrical terminals. The (complex) ratio of this force when the mechanical circuit is blocked (infinite impedance) to the current which produces it is the force factor. Since force is analogous to voltage (in the most com-

monly used system of analogues), the force factor is analogous to mutual impedance between two electrical circuits. It differs from the conventional electrical mutual impedance in that it makes no contribution to the electrical impedance when the mechanical circuit is blocked (secondary open-circuited) because its counter emf is due only to motion of the mechanical circuit, and in that the force factor in magnetic systems has opposite signs when viewed from the electrical and mechanical circuits. In usual circuit notation $z_{12} = -z_{21}$ ("antireciprocal" and not $z_{12} = z_{21}$ as in the electrical case). Since only the product of the force factors looking in both directions is involved in the following equations this will be called M^2.

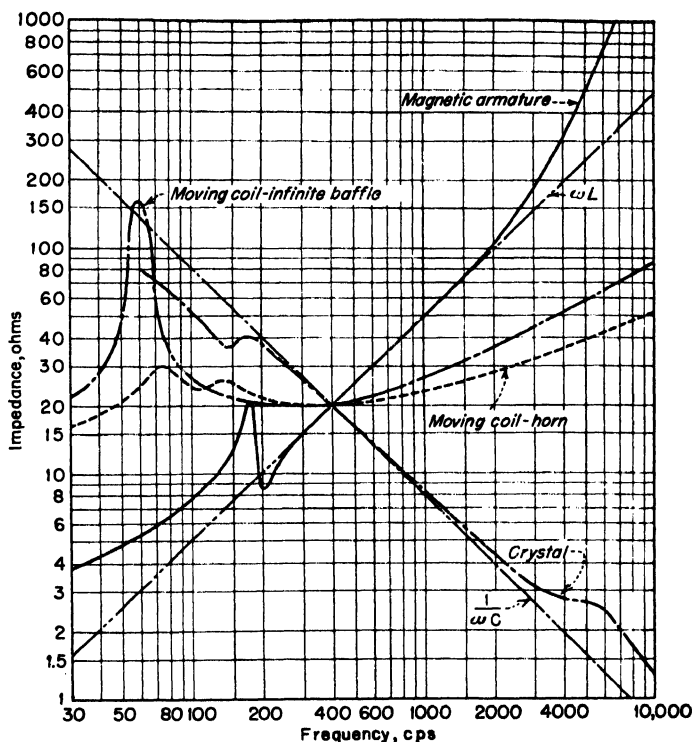



FIG. 27. Magnitude of normal input impedance of various speakers all adjusted to same value at 400 cycles to simplify comparison. Magnitude of reactance of pure capacitance and inductance shown for comparison.

The normal impedance of a speaker is defined as the impedance measured or seen at its signal terminals when operating normally with its proper acoustic load. The normal impedance z_N of moving-coil and magnetic-armature speakers is

$$z_N = z_e + \frac{M^2}{z_m} \tag{9}$$

where z_e = blocked electrical impedance of the speaker
z_m = total mechanical impedance seen by the mechanical circuit including diaphragm and acoustic load

In moving-coil speakers $M^2 = B^2l^2$, where B is the average radial flux density which the coil embraces and l is the conductor length. In balanced magnetic-armature

speakers $M^2 = 4B_0^2N^2/R^2$, where B_0 is the steady flux density in the gaps, N is the number of turns on the voice (armature) coil, and R is the effective reluctance of the alternating flux path (see Magnetic Armature).

In two-electrode electrostatic speakers

$$M^2 = -\left(\frac{E_0}{\omega d_0}\right)^2 \tag{10}$$

where E_0 = bias voltage (d.c.)
d_0 = biased electrode separation
$\omega = 2\pi f$

A two-terminal load impedance absorbs maximum power from a two-terminal source when the impedance of the load is the conjugate of the impedance measured or "seen" at the source terminals. The conjugate impedance is one having the same resistive or real part and a reactive or imaginary part equal in magnitude but opposite in sign. This holds for acoustical and mechanical circuits as well, but in these the terminals are not always so readily determined.

The speaker motor therefore absorbs maximum energy from the source, regardless of the complexity of the source network, when its normal impedance is the conjugate of the source impedance. The usual source is a vacuum tube or transistor amplifier which is approximately resistive. From Fig. 27 we see that the best match is a moving-coil or "dynamic" speaker.

FIG. 28. Sectional view of small moving-coil speaker showing structural simplicity.

16. Magnetic Motors. By the IRE definition, "a magnetic speaker is a loud speaker in which the mechanical forces result from magnetic reactions." This includes both moving-conductor or moving-coil (dynamic) and magnetic-armature speakers.

17. Moving Coil. A moving-coil motor is one in which the mechanical forces result from magnetic reactions between the field of the moving coil and the applied steady radial field in the air gap. A section of half of a moving-coil speaker is shown in Fig. 28.

Moving-coil motors are very widely used because (1) their electrical impedance permits good energy transfer from the source, (2) the large amplitudes required by the popular direct radiator diaphragms are obtained conveniently with minimum nonlinear distortion, (3) the mechanical impedance of the moving element may be made low, (4) the structure is simple and rugged mechanically, and (5) the cost is low.

The impedance seen at its electrical terminals when the coil is rigidly blocked is called the *blocked impedance* of the motor or speaker. This is approximately a high resistance and low inductance in series (R_e and L_e in Fig. 29) and is therefore easily coupled to an electronic amplifier. Near the fundamental resonance of the speaker the impedance rises, and, if a low-impedance source is used, the mismatch reduces the energy absorbed.

Moving-coil speakers are sometimes called *electrodynamic* or briefly *dynamic* speakers. Both terms have been applied for many years to speakers having either electromagnet ("energized") or permanent magnet fields. The prefix "electro" in electrodynamic has nothing to do with the source of steady flux in the gap.

18. Permanent and Electromagnets. Magnetic speakers require a source of magnetomotive force to provide the steady flux. If the current source is hum-free and

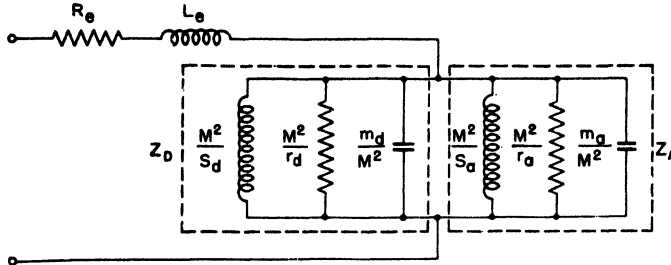



Fig. 29. Equivalent low-frequency electrical circuit of moving-coil or magnetic-armature speaker in a total enclosure or in an infinite baffle. In the latter case the enclosure stiffness S_a is zero and its equivalent electrical inductance infinite.

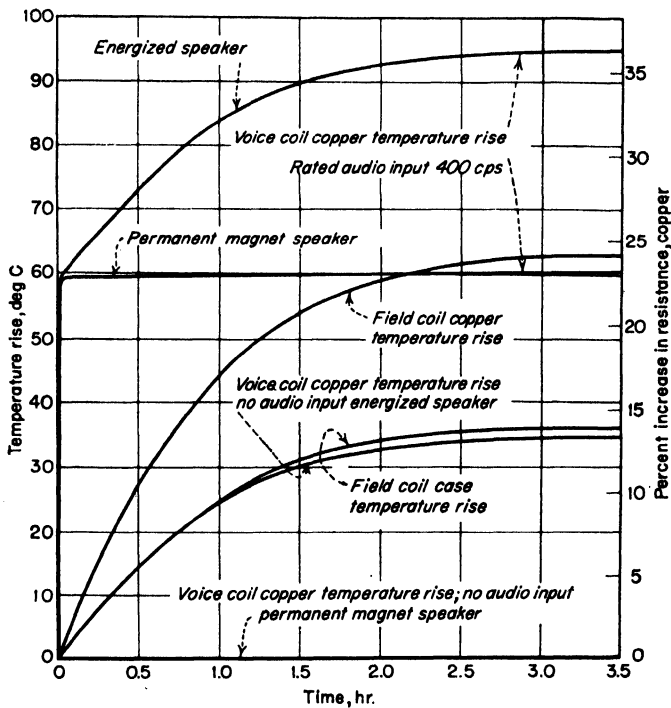



Fig. 30. Reduced voice-coil temperature rise in permanent-magnet speakers compared to energized types. Abnormal voice-coil temperature rise when rated complex-wave input is applied at 400 cycles also shown.

therefore the flux absolutely steady, the voice coil cannot distinguish between given flux densities due to permanent and electromagnets. The efficiency of any electromagnet speaker can be equaled or excelled by a permanent magnet if cost is neglected. In small motors the permanent-magnet type is usually less expensive. In intermediate-size motors the cost of the two types is comparable if the cost of a source of field power is included. In larger motors the permanent-magnet type may be more expensive though this is largely offset by the saving in power-supply cost. In

extensive speaker systems the installed cost of permanent-magnet types is frequently lower because of simplified low-cost wiring.

Permanent-magnet speakers are very widely used particularly when a special field current supply must otherwise be provided. The temperature of the electromagnet and consequently that of the voice coil rise with time as shown in Fig. 30. The field coil resistance rises, lowering the field current and flux density. The higher voice-coil resistance and reduced flux reduce the speaker efficiency. The higher voice-coil temperature reduces the permissible signal input power in voice-coil temperature-limited speakers.

The temperature rise when the rated complex-wave input is applied at a single frequency (400 cps) in a typical intermediate-size radio speaker is also shown. The single-frequency rating is normally much less than the "complex-wave" (speech and music) rating since in the latter case advantage is taken of the high ratio of peak to average power (normally 10 to 20 db; see Secs. 1, 2, 38).

19. Magnetic Armature. "A magnetic armature speaker (or motor) is a magnetic speaker (or motor) whose operation involves the vibration of the ferromagnetic circuit." The shorter term "magnetic" may be used where no confusion will result with moving-conductor or moving-coil speakers, which are also by definition magnetic speakers. A cross-sectional view of a balanced armature motor of this type is shown in Fig. 31. Flux increases in one pair of pole faces and decreases in the other pair, when current flows through the voice coil and when the armature moves, resulting in operation analogous to a push-pull circuit. The "voice" coil does not move and therefore is made relatively large. The resulting high inductance plus distributed capacitance in high-impedance types accounts for the large rise in impedance at high frequencies (see Fig. 27). This makes it difficult to couple it to an amplifier properly. To get high efficiency the armature pole-piece clearance must be small, and this leads to limited armature displacement and possible instability.

Fig. 31. Sectional view of balanced magnetic-armature speaker.

20. Electrostatic (Condenser) Speaker. An electrostatic (condenser) speaker (or motor) is a speaker (or motor) in which the mechanical forces result from electrostatic reactions. Structurally they are capacitors in which one electrode is free to move and serve as a diaphragm. There are two-electrode or "single-ended" and three-electrode or "push-pull" types. A d-c bias potential is applied between the moving and fixed electrodes to increase sensitivity and minimize distortion. The a-c potential is superimposed on the d.c. resulting in an alternating force on the diaphragm.

Some advantages of the electrostatic speaker are: (1) the diaphragm can be very light; (2) the force may be distributed uniformly or nonuniformly over the diaphragm; (3) this force distribution may be made frequency-dependent to control the directivity at different frequencies; (4) the transient and phase distortion may be kept low by proper design; (5) the cost may be low. Some of the disadvantages are: (1) the permissible diaphragm amplitude is normally low since, to attain reasonable efficiency, the electrode spacing cannot be large (at least without high bias potentials); this leads to large-area diaphragms or limits the low cutoff frequency; (2) large-area diaphragms tend to have poor directional properties and are normally either curved or warped to improve the directional properties, with consequent structural complications, or have the effective area modified by supplying potential to zones or areas on the diaphragm in a frequency-dependent manner by suitable

networks; (3) the blocked impedance is that of a capacitor (see Fig. 27) and even with the added motional impedance is difficult to match properly except in a limited frequency range; (4) a bias potential, which is normally substantially in excess of the a-c signal potential, must be supplied; (5) because of the high potentials dielectric problems are involved.

Two-electrode types with a flat circular diaphragm, a closely spaced rear electrode, and vents to control the impedance of the thin air film have been used for some time, largely in Europe, as inexpensive tweeters. Typical types have diaphragm diameters of 3.5 to 4 in. and a low-frequency cutoff of 3 to 4 kc. The upper

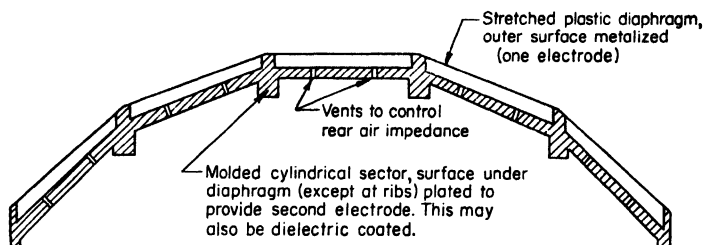



FIG. 32. Simplified cross-sectional view of a two-electrode electrostatic speaker of the type used to obtain an approximately cylindrical wave front at high frequencies.

cutoff may exceed 15 kc. Laboratory types are used up to 200 kc. For ultrasonic use, machining irregularities in the fixed electrode may leave sufficient air film to provide the high stiffness and yet the diaphragm displacement needed. Where a large-area h-f speaker (tweeter) is wanted without excessive beam concentration the construction shown in Fig. 32 or a group of rectangular elements may be used in a curved array.

To increase the permissible diaphragm amplitude with acceptable distortion and hence either increase the permissible power output or lower the low cutoff frequency, three-electrode push-pull or "double-ended" types are used. A schematic diagram of this type is shown in Fig. 33. When the impedance in series with the movable electrode is made high enough so that substantially uniform charge remains on the electrode throughout the a-c cycle, the permissible amplitude is markedly increased. The marked improvement which results has been investigated analytically and experimentally by Hunt.[1]

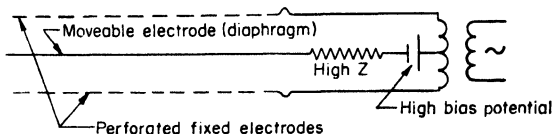



FIG. 33. Diagrammatic representation of a push-pull electrostatic speaker. If the impedance of the d-c bias source is very high, "constant-charge" low-distortion operation is obtained.

21. Crystal Speaker. "A crystal speaker (or motor) is a speaker (or motor) in which the mechanical forces result from the deformation of a crystal having converse piezoelectric properties." The crystal has a high mechanical impedance viewed from the driving point. Only a small displacement is possible without distortion or crystal fracture, so a mechanical transformer or lever arm is used when moderate excursions are required. This leads to mechanical complications, particularly at high frequencies where the transformer is not ideal, and to added cost. The application of this type has therefore been largely limited to h-f speakers in which the diaphragm amplitude is small. The blocked impedance is that of a leaky capacitance. The normal impedance of an 8-in. unit is given in Fig. 27.

[1] Hunt, F. V., "Electroacoustics," Wiley, 1954.

COMPLETE LOUD-SPEAKERS

The more important characteristics of a complete speaker system, which includes an electrical source of known impedance, one or more speaker units (combined motor and diaphragm), any dividing networks (filters), and specified acoustic load, are its pressure response, efficiency-frequency, impedance, directional, and distortion characteristics (see Tests).

22. Pressure Response. The pressure response may be defined as the sound pressure, in dynes per square centimeter, at a designated point in the medium with respect to the speaker, per volt input. This quantity is very useful when the speaker is to be operated from a very low impedance or constant-voltage source and listened to outdoors at the same point since these conditions correspond to the test conditions. If the source (amplifier) voltage is not constant, the source is connected to the speaker, under normal use conditions, and the speaker input voltage vs. frequency is determined as the frequency is varied either by applying a constant input voltage to the amplifier or, for a complete system, by applying constant sound pressure to the input microphone.

If the speaker is reasonably free of nonlinear distortion the sum of this curve and a pressure response vs. frequency curve for the speaker will give the combined amplifier speaker frequency response.

23. Reference Pressure Response. A definition of this term which has been proposed by the IRE and ASA follows: "The reference pressure response of a loudspeaker at a specified frequency expressed in db is the difference between the output sound pressure level in a free field on the principal axis referred to a (*i.e.*, computed for a) distance of one meter, and the input electrical power level delivered to the loudspeaker rating impedance." "The rating impedance is the value of a pure resistance which is to be substituted for the loudspeaker when setting the electric power supplied by the source." The rating impedance is specified by the speaker manufacturer and is the value he decides gives the best average "match" to the source when the conflicting factors of distortion, frequency response, etc., are considered. The reference levels are 0.0002 dyne per sq cm (microbars) for pressure p_0 and 1 mw for electrical power P_0. That is, the reference pressure response G_L is given by

$$G_L = L_L - L_E = 20 \log_{10} \frac{p_L}{p_0} - 10 \log_{10} \frac{P_E}{P_0} \tag{11}$$

This penalizes the speaker if its electrical input impedance differs from the "rating impedance," which is really a rating resistance.

24. Efficiency-frequency Characteristic. The energy efficiency, or simply efficiency, of a loud-speaker is the ratio of the useful acoustic-energy output to the signal-energy input. A definition for the "rated efficiency" G_{LE} proposed by the IRE and ASA is

$$G_{LE} = 10 \log_{10} \frac{P_L}{P_E} \tag{12}$$

where P_L is the total radiated acoustic power in watts, P_E is the electric power delivered to the "loud-speaker rating impedance" in watts, and

$$P_E = \frac{E_G^2 R_{LR}}{(R_{LG} + R_{LR})^2} \tag{13}$$

where E_G is the generator (amplifier or source) voltage in volts, R_{LR} is the loud-speaker rating impedance (resistance) in ohms, and R_{LG} is the "loud-speaker measurement source impedance" (resistance) in ohms defined as "the value of a pure resistance to be connected in series with the speaker and a constant voltage source in order to measure the speaker performance."

The rated (sometimes called the "available power" or "system") efficiency penalizes the speaker if its electrical input impedance differs from the "measurement

source resistance" R_{LG}. The rated efficiency and the efficiency (conventional or "energy") as defined above may differ by a factor of 10 or more in a direct radiator speaker at the low fundamental resonant frequency if the force factor is high and the mechanical resistance low since in this case the l-f peak in the impedance may be much higher than shown in Fig. 27.

If the effective internal resistance of the source or "measurement source impedance" R_{LG} equals the loud-speaker rating impedance R_{LR}, then the rated efficiency is alternatively given in terms of the speaker parameters by

$$G_{LE} = \left| \frac{4R_{LR}M^2r_r}{[z_e + (M^2/z_m)]^2 z_m^2} \right| 100\% \tag{14}$$

where z_e = blocked voice-coil impedance plus r_s
z_m = total mechanical impedance of the mechanical mesh including diaphragm radiation and air load
r_r = total radiation resistance seen by diaphragm
M is defined under Force Factor. The vertical lines indicate that the absolute value is to be taken.

In a direct-radiator moving-coil loud-speaker mounted in an infinite baffle the rated efficiency is not constant with frequency. However, there is one region, the piston range, in which a fortuitous combination of characteristics yields a reasonably constant efficiency. For the ordinary large direct-radiator speaker this usually occurs over a substantial portion of the 100- to 800-cps range. In smaller sizes this range may be limited to 400 to 1,000 cps. Here most cones vibrate as rigid pistons, so that their radiation is closely that of a piston of the same projected area. Next, $2\pi R/\lambda$ is less than unity, so that the radiation resistance r_r varies with the square of the frequency (see Fig. 11). The frequency at which the motor is resonant usually lies well below that at which $2\pi R/\lambda$ is unity, and since the diaphragm is pistonlike, the mechanical reactance in the piston range is predominantly masslike. Also, the impedance of the voice coil is largely resistive and nearly constant. If it is coupled to a source of equal resistance and if radiation from only one side of the diaphragm is considered, it turns out that the piston-range system efficiency is equal to about $5 \times 10^{-11}\alpha E d^4/m^2$. Here α is the ratio of the volume of the voice-coil conductor (assumed to be copper) to the total gap volume, E is the magnetic energy in the gap, in ergs, d is the effective diameter of the diaphragm, in centimeters, and m is the sum of the mechanical and air masses of the moving system, in grams. From this expression it is seen that the contribution of the magnetic structure to the speaker efficiency is in terms of the gap energy, rather than the flux density or the total flux in the gap. Since the gap energy is roughly proportional to the magnet weight or the field-coil copper weight, it is seen that economic factors play a large part in determining the efficiency, at least in the piston range.

The 400-cps rated efficiency of the speakers commonly used in radio receivers ranges from 1 to 4 per cent. The corresponding 400-cps efficiency of direct-radiator speakers with very large motors ranges from 5 to 20 per cent. The rated efficiency near the low resonant frequency will be substantially less because of the poor impedance match to the source. Efficiencies of this order are more readily obtained in horn speakers, but 30 per cent is rarely exceeded over an extended frequency range. Although higher values are frequently claimed, these values, if based on any measurements, are usually based on motional impedance measurements in which all horn, diaphragm, air, eddy-current, and hysteresis losses have been assumed to be useful acoustic radiation.

25. Baffles, Enclosures, and Cabinets. "A baffle is a partition which may be used with an acoustic radiator to increase the effective length of the acoustic transmission path between front and back of the radiator." This term is usually reserved for a relatively flat baffle in which both sides of the diaphragm look into substantially a hemisphere (solid angle of 2π steradians). The term *directional baffle* is sometimes

applied when one side of the diaphragm looks into a smaller solid angle. The baffle then begins to take on the properties of a horn. There is no sharp line of demarcation, but there appears to be little reason for calling any structure which restricts the solid angle to less than $\pi/2$ (an octant of a sphere) anything but a horn.

If a baffle is used outdoors, appreciable destructive interference or pressure cancellation between the front and back waves of the speaker may occur at the listener's position at some frequency above the cutoff frequency. The frequency at which this occurs depends on the baffle size and listener location. Destructive interference at the cone itself is usually unimportant except near the cutoff frequency of the baffle. To distribute this effect and make it cover a broad band, baffles shaped as shown in Fig. 34 are sometimes used. Since the major effect depends primarily on the listener's location and a lack of reflected waves, no such simple result occurs indoors and a space average of the pressure in a moderate-size listening room shows no such effect. Conventional rectangular baffles may therefore be used indoors unless the room approaches free-field or outdoor characteristics.

FIG. 34. Irregular baffle shapes used outdoors to broaden frequency band of destructive interference between speaker front and back waves at listener's position.

The equivalent l-f electrical circuit of a moving-coil or magnetic-armature speaker in an infinite baffle is shown in Fig. 29. Here R_e and L_e are the blocked voice-coil resistance and inductance. Z_D is the electrical equivalent of the diaphragm less air load. Z_A is the equivalent of the air load, except that in this case there is no stiffness S_a provided by the air load, so its equivalent inductance M^2/S_a is infinite. M is defined in Sec. 15. S_d, m_d, and r_d are the effective diaphragm stiffness, mass, and resistance, respectively, and r_a and m_a are the radiation resistance and mass which may be determined from Fig. 11. Note that both sides of the diaphragm have radiation resistance and mass in this case and the values *per unit area* given in Fig. 11 must be multiplied by twice the piston area to give the r_a and m_a used in Fig. 29.

The magnitude of the impedance of a moving-coil speaker in an infinite baffle is shown in Fig. 27. The antiresonant impedance of the parallel circuit corresponds to the resonant frequency of the diaphragm and air load and is limited by the parallel value of the two resistances. These resistances are proportional to the square of the flux density and inversely proportional to the diaphragm and air (radiation plus dissipation) resistances. A high resonant impedance is therefore not necessarily undesirable, as is generally supposed, since it may be due to a high flux density and therefore mean a high efficiency over a wide frequency band.

FIG. 35. One type of vented enclosure in which port area is large and placed near diaphragm to obtain maximum aid from mutual radiation resistance between diaphragm and port. Phase shift of backside radiation is obtained by choice of circuit constants.

The effect of the source impedance, connected to the input terminals, on response may be noted from this circuit. If the source resistance R_{LG} is low, the speaker will absorb very little power at resonance and the acoustic output may not rise appreciably. The voice-coil resistance R_e and the source resistance R_{LG} in series are effectively across the antiresonant circuit at low frequencies since the reactance of L_e can be neglected. When the flux density is high, the Q of the antiresonant circuit alone is high but the

source and voice-coil resistances then provide substantial shunt resistance and "electromagnetic damping." The effect of this on the speaker response to a pulse is shown in the experimentally determined curves of Fig. 50. The minor irregularities in these damped sinusoids are due to h-f modes of vibration of the diaphragm. As the source resistance is raised, relatively more power is supplied the speaker at antiresonance and at high frequencies where the effect of the voice-coil inductance is important and less is supplied in the mid-frequency range. The steady-state response of a vented enclosure (see Figs. 35, 36, 37) as the source resistance or "impedance match" is changed is shown in Fig. 49. Here the response at 400 cps has arbitrarily been adjusted to the same value as the source resistance was changed.

Fig. 36. Equivalent low-frequency electric circuit of moving-coil or magnetic-armature speaker in vented enclosure.

26. Total Enclosure. A total enclosure which prevents radiation from the back side of a diaphragm may be used to prevent destructive interference between the front and back waves from a diaphragm. This might be obtained by closing the vent or port in Fig. 35. When the wavelength exceeds four times the maximum enclosure dimension, the enclosure adds a *total* stiffness S_a viewed from the diaphragm, $S_a = \rho c^2 A_d^2 / V_0$ cm per dyne, where A_d is the effective piston area of the cone in square centimeters and V_0 is the net volume of the box in cubic centimeters. The

Fig. 37. Impedance of circuit of Fig. 36.

"compliance" is the reciprocal of this value. This stiffness raises the natural frequency of the speaker. If the enclosure includes absorbing material, this stiffness will be altered by the reactance seen at the surface of the material. Each square centimeter will dissipate $P^2 \times 10^{-7}/r_B$ watt, where P is the sound pressure in the box and r_B is the resistance per unit area for sound of normal incidence on the absorb-

ing material. From this the equivalent resistance in parallel with the box stiffness may be obtained.

The l-f equivalent electrical current of such an enclosure is shown in Fig. 29. Here Z_D is the electrical equivalent of the diaphragm alone; S_d, r_d, and m_d are the stiffness, resistance, and mass of the diaphragm measured *in vacuo*. The electrical equivalent of the air load including radiation impedance is Z_A; S_a is the effective enclosure stiffness, r_a the total air or fluid resistance (enclosure and grill cloth or opening resistance if any, and radiation resistance), and m_a is the effective air (radiation-plus-enclosure) mass. Normally the parallel value of Z_D and Z_A or a single parallel "antiresonant" circuit is shown but the contributions of individual elements are then not as clear. Since the electrical circuit elements are *inverses* with respect to M^2 [see Eq. (9)], the stiffness appears as an inductance and the mass as a capacitance.

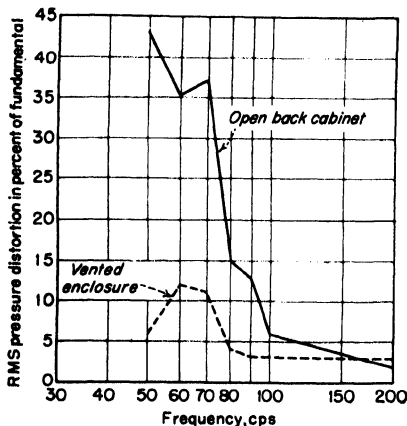



Fig. 38. Total distortion of open-back cabinet and of the same cabinet enclosed as shown in Fig. 35; same speaker and electrical input in both cases. Low-frequency distortion is much reduced because diaphragm sees high antiresonant impedance of enclosure and therefore has only small displacement whereas air in port (which lacks the nonlinear edge stiffness and nonuniform flux of the speaker) moves with large displacement.

If the volume is small enough or the natural frequency of the speaker out of the enclosure low enough, the enclosure and not the diaphragm stiffness will control the natural frequency.

There is little to gain in making the mechanical stiffness of the speaker less than one-fifth that of the enclosure S_a given above, since this mechanical stiffness raises the resonant frequency by only 10 per cent and it permits using a good mechanical suspension and good centering, thereby economizing on magnet structure volume and cost. Figure 39 shows what the resonant frequency of an 8-, 12-, and special 15-in. speaker (designating sizes) would be as a function of *net* enclosure volume (gross volume less volume of speaker cone and magnet structure) if the speakers had *no* mechanical stiffness. A typical 1-in. voice-coil 8-in., a typical 1.5-in. voice-coil 12-in., and a special 15-in. "woofer" were assumed. The 15-in. has an 80-g l-f dynamic mass, including radiation or air mass, because of the heavy voice coil and cone used to optimize its performance as an l-f speaker. The 12-in. speaker has a dynamic mass of only 27 g but because of its reduced diaphragm area sees only one-third as much acoustical or air stiffness in the same enclosure, hence resonates at the same frequency as the 15-in. speaker. The 8-in. diaphragm assembly has a mass of only 6.8 g but resonates at a lower frequency because of the lower acoustic stiffness it sees.

From Fig. 39 note that an enclosure of say 1.5 cu ft would result in a resonant frequency of 96 cps for either the standard 12-in. or special 15-in. speaker even if they had no mechanical stiffness. If the enclosure volume is kept at 1.5 cu ft the cost in dollars and/or efficiency of lowering the resonant frequency is high. If we halve the frequency to 48 cps by keeping the same voice coil and magnet but increase the dynamic mass by a factor of 4, by using a heavier cone or otherwise, the efficiency in the useful band drops 12 db, or to one-sixteenth its original value above 96 cps. The efficiencies of the modified and unmodified speakers converge at 46 cps. The original efficiency cannot be maintained in practice with the lower resonance because the energy in the air gap cannot be raised by the required factor of 16 (when all the mass increase is in voice-coil conductor) without prohibitive cost and magnet size increases.

The compromises between enclosure size, radiated power, and resonant frequency can be kept in mind by reviewing the following factors. The stiffness seen by the diaphragm is proportional to the square of the diaphragm area. The efficiency in the pass band drops 12 db for each octave the resonant frequency is lowered, corresponding to 6 db for a 29 per cent and 2 db for a 10 per cent lowering of the resonant frequency. If the diaphragm is driven at constant amplitude, the radiated power varies with the square of the diaphragm area.

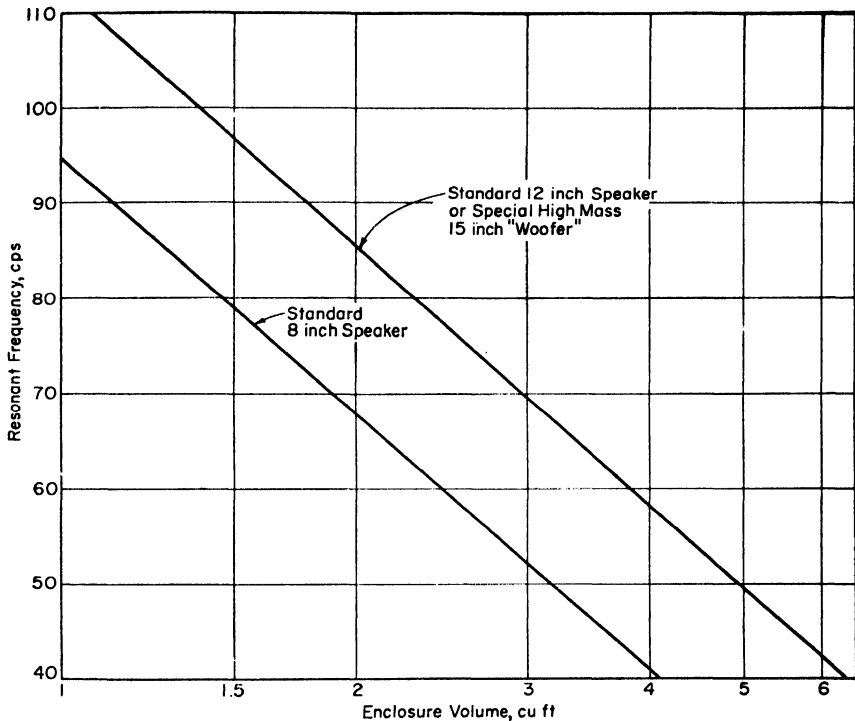



FIG. 39. Resonant frequency of standard 8-in., 12-in., and a special 15-in. speaker as a function of net enclosure volume for the idealized case of no diaphragm mechanical support stiffness.

In practice, therefore, the resonant frequency should not be lowered needlessly. With good music listening in mind 60 to 70 cps is adequate. Lower frequencies are of primary interest to the experimenter and "acoustic effects" man. If the enclosure must be small, the diaphragm may be small to lower the resonance but the radiator power should be correspondingly reduced. The most satisfactory way of keeping reasonable efficiency and l-f response is to use an adequate enclosure volume. A minimum of several cubic feet is desirable. If greater l-f output is needed, a properly designed vented enclosure should be considered.

A total enclosure is sometimes called an *infinite* baffle. While it resembles one in preventing front and back wave interference, it has two important differences which make this designating term undesirable. The enclosure adds an air impedance to the rear of the diaphragm, which may be very different from that seen in an infinite baffle. An infinite baffle restricts the radiation to a hemisphere, and the radiation impedance seen by the diaphragm is given by Fig. 11. If the enclosure is used outdoors, the radiation resistance which the outside of the diaphragm sees at low fre-

quencies is only half this value and the reactive part approximately seven-tenths this value. In practice the useful efficiency is almost halved at low frequencies. Indoors the impedance seen will depend on the environment as described in Sec. 5, also in Room Acoustics, below.

27. Vented Enclosures. The idea of putting a vent or "port" in an enclosure is very old. It was first done to provide "pressure relief." In more recent types, known as "vented" enclosures, detailed consideration has been given to the very important effect of the mutual impedance between the port and diaphragm. The port area is large and the port is near the diaphragm to increase the mutual radiation resistance and extend the frequency range over which it is effective (see Sec. 7). Such an enclosure is shown in Fig. 35. The effective or virtual diaphragm in the opening is coupled internally through the stiffness of the air in the enclosure to the diaphragm. It is also coupled externally. The equivalent l-f circuit is shown in Fig. 36. Impedance Z_D corresponds to Fig. 29 and r_a and m_a correspond except that the mutual-radiation impedance must be added.

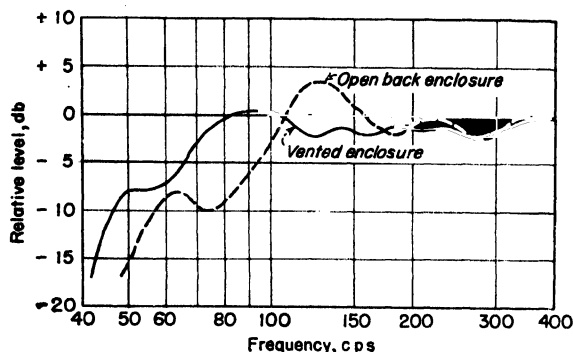



FIG. 40. Total sound power radiated by cabinet for which distortion is shown in Fig. 38. Shape of low-frequency response may be varied between wide limits depending on enclosure volume, port area, and speaker used.

The vent and enclosure have therefore added one *LRC* circuit. The effect of this is to shift the back-side-cone radiation by nearly 180 deg above the frequency at which the port mass m_p and box stiffness *viewed from the port* are resonant when the cone is blocked. For about one-third of an octave above and below this frequency most energy is radiated by the port. Although the diaphragm and port radiation are out of phase below this frequency, the port radiation greatly exceeds the diaphragm radiation near this frequency.

The enclosure is made as compact as possible. The port can be placed near the diaphragm to increase the mutual-radiation resistance since the phase shift is not primarily due to transmission time delay but occurs because the acoustic circuit goes through antiresonance, the phase shift occurring suddenly at this frequency. In properly designed enclosures, advantage is taken of a large mutual-radiation resistance to improve the l-f efficiency. Very little absorption in the enclosure is wanted at low frequencies to take maximum advantage of backside radiation. At frequencies of several hundred cycles or more where the port radiates negligible sound the enclosure is made absorbent to avoid "box" resonance. The advantages of vented enclosures are: (1) backside radiation is used to increase substantially the l-f output; (2) most of this output comes from the port which has no nonlinear diaphragm suspension stiffness to produce nonlinear distortion; (3) antiresonance of the enclosure occurs near the lower frequency of maximum radiation so the diaphragm amplitude is much less than it would be otherwise. The result of these factors on nonlinear distortion reduction is shown in Fig. 38, in which the effect of converting an open-back cabinet to a vented port enclosure of the same internal volume is shown. The change in response is shown in Fig. 40.

While a large number of vented enclosures, which differ in physical appearance, have been made they all have substantially the equivalent circuit shown in Fig. 36. The inclusion of absorbing material may introduce a resistance in the box stiffness branch, and various values of the parameters are used to modify the response. In

some instances, the relative increase in l-f response is actually obtained by reducing the middle-frequency response.

28. Transmission-line Speaker. The phase and amplitude of the backside radiation of a cone may be altered by coupling a conduit or acoustic transmission line to it. In early types the multiple-resonant properties of such a line were used to influence the response. In a more recent type, known as the labyrinth, the line is folded to conserve space and made highly dissipative (see Fig. 41). Phase shift between the diaphragm and port or open end is due to time of transmission in the line. At very low frequencies the line is a small fraction of a wavelength long, the phase shift is negligible, and the port and diaphragm radiation are out of phase. When the line is $\lambda/4$ long, it acts as an impedance inverter (as in the electrical case); the cone sees a high impedance, and the radiation from the port is a maximum. Nonlinear distortion is therefore reduced at and near this frequency. The resonant frequency of the diaphragm may be placed at this frequency to aid damping. Between this frequency and the one for which the line is $\lambda/2$ long, the port phase shifts gradually but maintains some component of its radiation in phase with the diaphragm (neglecting separation between the port and diaphragm) outside the line. Because of the infinite series of resonant and antiresonant frequencies of the line, high absorption must be introduced to prevent the production of objectionable resonances and radiated out-of-phase components of the port. Most of the rear-side radiation is therefore absorbed. The comparison of the responses of an open-back cabinet and labyrinth is shown in Fig. 42.

FIG. 41. Labyrinth type of transmission-line speaker. Phase shift of backside radiation is obtained by time of transmission delay in line.

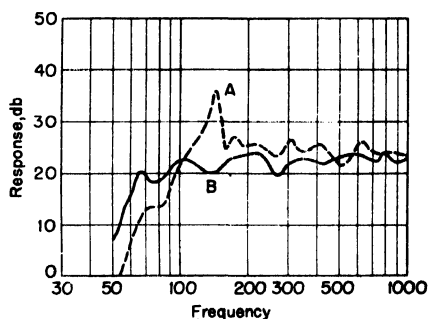



FIG. 42. Relative response of open-back cabinet A and labyrinth B.

ROOM ACOUSTICS

29. Room Characteristics. The trend in the theory of room acoustics is toward considering the source of sound, the room, and the sound receiver or "sink" all as part of a unified dynamical system. This is required to bring out the interaction between source, sink, and room and their effects on the steady-state and transient aspects of sound transmission in the room.

In this theory the room is considered as an assemblage of resonators and the walls of the room as terminal impedances determining absorption and reflection. A rectangular room has a triple infinity of resonant frequencies. If the wall impedances are high pure resistances, these frequencies are given by

$$f = 17{,}140\left[\left(\frac{n_x}{l_x}\right)^2 + \left(\frac{n_y}{l_y}\right)^2 + \left(\frac{n_z}{l_z}\right)^2\right]^{1/2} \tag{15}$$

where $n_x, n_y, n_z = 0, 1, 2, \ldots$

l_x, l_y, l_z = dimensions of rectangular room, cm

The distribution of these "allowed" frequencies (at which resonance occurs) may be graphically shown as in Fig. 43 by a three-dimensional plot in "frequency space." Each vector to a lattice point is associated with a "natural frequency" or "normal mode" of the room. The shortest vector, corresponding to the lowest frequency, is determined by the longest dimension of the room. The direction of the vector from the origin to a lattice point indicates the direction of excitation of that frequency in the room, and the length of the vector is proportional to its frequency.

At low frequencies there may be an appreciable frequency interval between the natural frequencies if the room is small. At high frequencies the number of natural frequencies in a given frequency interval is proportional to the square of the frequency.

FIG. 43. Distribution of resonant frequencies for a rectangular room with side lengths l_x, l_y, and l_z. The length of a vector from the origin to each lattice point indicates the frequency and the direction of the vector indicates the direction of the corresponding standing wave. The velocity of the sound wave c is given by Eq. (1).

30. Reverberation. Using this concept of multiple natural frequencies, the decay of sound in a room may be described as follows: Assume energy has been supplied to the room until the energy level is constant; *i.e.*, the rate of absorption at the boundaries equals the rate of supply to the room. The resulting standing-wave system depends not only on the room and frequency but on the location and orientation of the source. When the source of energy is stopped each individual mode of vibration of the room will decay exponentially at its own frequency and the combined effect of these is called *reverberation*. Only the modes having allowed frequencies near the fre- contain appreciable energy. quency of the steady-state excitation will

By definition the reverberation time is the time required for the mean energy density in the room to drop 60 db. While this mean may be the result of a large number of rates of decay each of which is individually exponential, the combined value in general is not given by a single exponential term. This accounts for the fact that the slope of the mean-energy-density time-decay curves for the average room are not uniform and therefore for the fact that the apparent reverberation time depends on the time interval over which the decay is averaged. If the absorption is moderate the approximate reverberation time in seconds is given by

$$T = \frac{0.00161V}{-A \ln (1 - \alpha)} \tag{16}$$

where V = room volume, cc

$\alpha = (A_1\alpha_1 + A_2\alpha_2 + \cdots)/A$ total room absorption

A_1, A_2, etc. = areas, sq cm, having absorption coefficients α_1, α_2, etc., respectively

A = total boundary area of room

Corresponding to this type of energy decay in the room, there is a growth curve. When a source suddenly emits energy, each of the excited modes absorbs energy in an exponential manner. This occurs until the asymptotic or steady-state value is reached after an infinite length of time. When this is reached, the acoustic power supplied to the room must equal that absorbed at the room boundaries. By definition the absorption coefficient α of the boundary is the fraction of the incident energy absorbed for a specified angle of incidence. The intensity I of a sound wave [Eq. (2)] is a measure of the energy per square centimeter per second. The total intensity is made

up of two parts: that due to the reverberant sound and that due to sound received directly from the source. The acoustic power radiated by the source is related to the total intensity by

$$P_a = \frac{I}{\frac{1-\alpha}{A\alpha} + \frac{1}{4\pi r^2}} \quad \text{watts} \tag{17}$$

where P_a = radiated power, watts
I = total intensity, watts per sq cm
r = distance between source and point at which I is measured, cm

Table 3. Representative Acoustical Coefficients

Manufacturer	Material	Thickness, in.	Absorption coefficients, Frequency, cps: 125	250	500	1,000	2,000	4,000
Armstrong Cork Co.	Cushiontone A-1	½	0.03	0.14	0.56	0.82	0.79	0.59
Armstrong Cork Co.	Corkoustic B-5	1¼	0.03	0.12	0.52	0.62	0.41	0.46
Armstrong Cork Co.	Arrestone	2½	0.34	0.69	0.99	0.94	0.74	0.59
Celotex Corp.	Acousti-Celotex C-1	½	0.05	0.15	0.64	0.87	0.76	0.78
Celotex Corp.	Celotone	¾	0.09	0.27	0.83	0.92	0.77	0.72
Celotex Corp.	Acousteel	1¼	0.35	0.71	0.99	0.90	0.70	0.55
Johns-Manville Sales Corp.	Fibretone	½	0.03	0.21	0.62	0.68	0.72	0.68
Johns-Manville Sales Corp.	Fibretone	¾	0.10	0.32	0.81	0.90	0.79	0.67
Johns-Manville Sales Corp.	Sanacoustic	...	0.29	0.68	0.99	0.99	0.89	0.52
National Gypsum Co.	Econacoustic	½	0.03	0.13	0.67	0.77	0.72	0.72
National Gypsum Co.	Travacoustic	¾	0.08	0.23	0.87	0.89	0.90	0.85
National Gypsum Co.	Acoustimetal	1¼	0.26	0.62	0.99	0.99	0.83	0.59
Owens-Corning Fiberglas Corp.	Fiberglas acoustical tile type TMW	¾	0.03	0.13	0.69	0.95	0.94	0.85
United States Gypsum Co.	Random perforated auditone	½	0.11	0.26	0.52	0.57	0.60	0.66
	Acoustone	¾	0.15	0.29	0.84	0.93	0.78	0.71

Description of object	128	512	2,048
Balcony opening; soft seats	.....	0.50-1.00	
Brick wall, unpainted	0.024	0.03	0.049
Carpet, no pad	0.09	0.20	0.27
Carpet, felt pad	0.11	0.37	0.27
Draperies, 10 oz per sq yd	0.04	0.11	0.30
Draperies, 18 oz per sq yd	0.10	0.50	0.82
Floors, concrete	0.01	0.015	0.02
Floors, linoleum	.....	0.03-0.08	
Glass	0.035	0.027	0.02
Plaster, smooth, on tile	0.013	0.025	0.04
Wood paneling	0.08	0.06	0.06

Description of object	Total absorption, equivalent area, sq cm		
Audience, seated, per person	1,000 2,000	3,000-4,000	3,500-6,000
Chairs, metal or wood	150	170	200
Chairs, theater, heavily upholstered		2,500-3,000	

With the usual placement of listeners in studios and theaters the contribution of the second term is small in the audience space, and most of the intensity is due to the reverberant sound. The foregoing assumes that I is uniform throughout the room. This is reasonably valid if the room has a reverberation time near optimum, if the sound source is not highly directional, if the room is many wavelengths long, if its proportions are neither too cubical nor too elongated, and if large flat surfaces are broken up by diffusing shapes or patches of absorbing material.

When the direct sound intensity is relatively small, Eqs. (16) and (17) may be combined to yield

$$I = \frac{P_a e^{-0.00161V/TA}}{A(1 - e^{-0.00161V/TA})} \quad \text{watts/sq cm} \qquad (18)$$

This relation may be used to obtain the efficiency of a loud-speaker by measuring the intensity produced, calculating the acoustic output from Eq. (18), and comparing that with the electrical power available to the speaker from the electrical source.

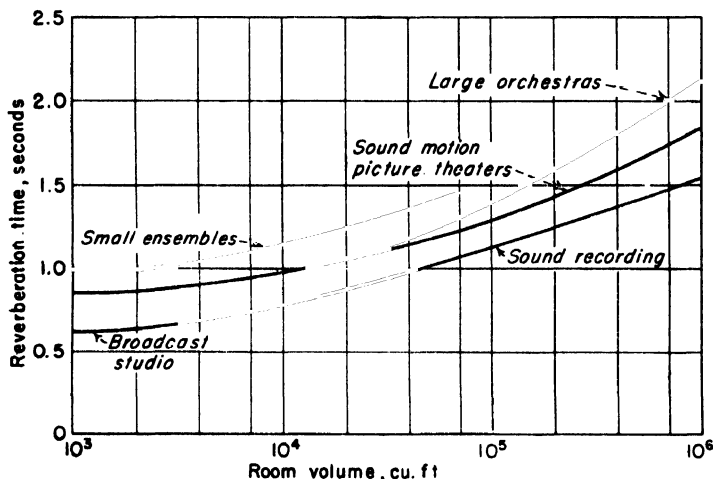



FIG. 44. Dependence of optimum reverberation time on room volume for various applications. Various authorities differ by 20 per cent or more from the indicated values so this deviation may be considered permissible.

In Table 3 are listed the absorption coefficients for some representative proprietary acoustical materials and objects, and the total absorption for persons and seats, for use in Eq. (12).

31. Room Power Requirements. If we know the desired sound intensity, the acoustic input power P_a required to produce it may be obtained either from Eq. (17) by knowing the total room absorption or from Eq. (18) by knowing the room volume and reverberation time. If the shape of the room is unknown, the surface area A may be approximated by $SV^{2/3}$, where S is a factor between 6 for a cubic room and about 12 for an elongated room. Desirable reverberation times in terms of room volume for various uses are shown in Fig. 44. The values for speech only will lie near or below the lowest curve and are seldom realized except in acoustically treated rooms. Typical schoolrooms with average attendance, for example, usually have reverberation times well up toward the upper music range.

Speech articulation increases rapidly with intensity up to an intensity level of 40 db or 10^{-12} watt per sq cm and more slowly to 50 db or 10^{-11} watt per sq cm. If room noises are present, the speech intensity should exceed these by at least 10 db and preferably by 20 db or more. In conversational speech the person speaking

radiates about 10^{-5} watt. Loud speaking requires 10^{-3} watt. If a loud-speaker is to simulate a person speaking loudly, its acoustic output should be at least 10^{-3} watt, which, for a 1 per cent efficient loud-speaker radiating all its output into the room, means an electrical input of $\frac{1}{10}$ watt.

There is considerable difference of opinion on what constitutes acceptable levels of reproduced sound. Values of electrical power input which have been suggested for theater use are shown in Fig. 45. These are based on the use of speaker systems with average system efficiencies of 25 per cent. The trend is toward larger inputs to get enhanced dramatic value in the reproduction.

In Fig. 46 are shown the *minimum* recommended acoustical power inputs to rooms having acoustical characteristics suitable for sound motion-picture theaters, for a sound pressure level of 80 db. This is satisfactory for speech in a fairly quiet room.

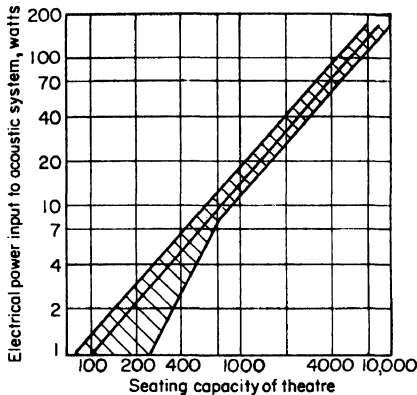



FIG. 45. Recommended amplifier output for motion-picture reproduction. Speaker system efficiency assumed to be 25 per cent; so acoustic input to theater is one-quarter indicated electrical input. Trend is toward higher values.

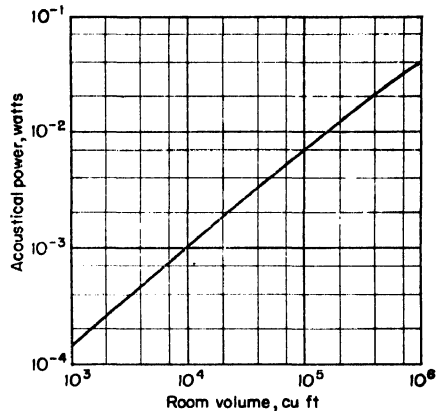



FIG. 46. Acoustical power to produce 80-db level in rooms acoustically similar to sound motion-picture theaters.

For music reproduction, various authorities recommend inputs some 15 to 25 db higher to simulate the actual orchestra. This increases the amplifier power needed some 30 to 300 times above the recommended minimum for speech. From this it may be seen that differences of opinion may account for large performance and cost differences.

32. Acoustic System Characteristics. When a sound receiver is included in the room with a source, we must consider the reaction of the room on it. The most common receiver is a listener. Because of the difficulty, however, of making objective measurements of what is going on in the listener's central nervous system, it is more convenient, although only approximately correct, to substitute one or more microphones for the listener.

For sake of simplicity, assume we have a loud-speaker as a source and a microphone as a receiver or sink, each with two accessible terminals. Since no source of energy is assumed in the room, these four leads may be considered the terminals of a passive quadripole or four-terminal network. From circuit theory we know that the measurement of three independent quantities will completely specify the performance of this quadripole at any one frequency. By analogy with the purely electrical case we may define the ratio (complex) of the current through a specified microphone load to the input voltage of the speaker as a transfer admittance.

A curve giving the magnitude of this quantity for a speaker and a microphone

mounted in diagonally opposite corners of a rectangular 18- by 20- by 11-ft room is shown in Fig. 47. This transfer admittance is what might be called the response curve of the loud-speaker measured in this room with designated locations for the source and microphone and with the particular microphone employed. As would be expected from the complicated equivalent circuit of the room, this transfer admittance varies by a large factor with frequency and exhibits a large number of maximums and minimums at high frequencies. In passing it may be noted that the impedance of the boundary of this room varied appreciably with frequency and therefore the locations of the resonant and antiresonant frequencies of the "electroacoustoelectrical network" do not occur at the frequencies predicted by constant boundary impedances.

Considering the matter in this light, we see that this response curve depends on the type of speaker, microphone, their location in the room, the geometry of

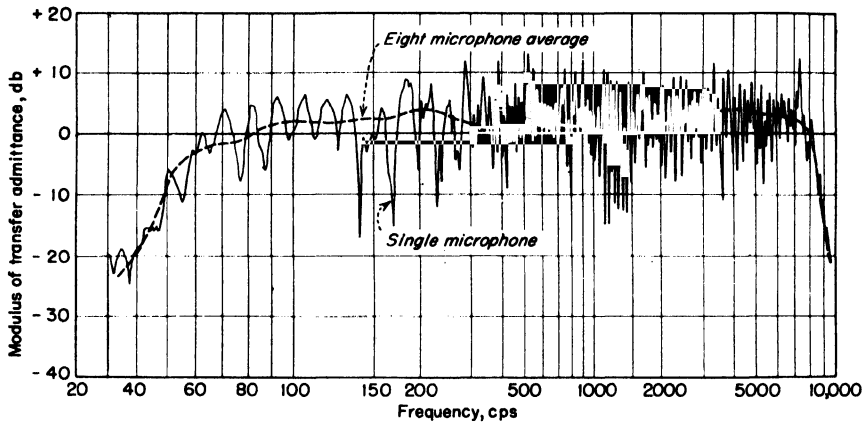



Fig. 47. Transfer admittance or "response curve" of a speaker, room, and microphone or electroacoustoelectrical transducer. Speaker and microphone are in diagonally opposite corners of an 18- by 20- by 11-ft room.

the room, and the impedance of the room boundaries and therefore the impedance of the entire dynamical system seen from the acoustic terminals of the speaker and microphone.

The loud-speaker supplies maximum energy to the room when the impedance seen by the diaphragm looking into the room is the conjugate of the impedance looking back into the diaphragm with the speaker connected to its generator or amplifier. Since the latter impedance is usually high, the speaker supplies maximum power when the room impedance is high, which occurs (by definition) when the ratio of the pressure to particle velocity is high, *i.e.*, when the speaker is near a pressure maximum. A pressure-actuated microphone gives maximum response at a pressure maximum. The maximum peaks in the transfer admittance of Fig. 47 therefore occur when both the microphone and speaker are near pressure maxima. No resonance pressure maximum occurs at the speaker below the lowest resonant frequency of the room, and good l-f response is therefore hard to obtain in small rooms.

The apparent damping or Q of a mode of the system may be obtained by steady-state sharpness of resonance methods corresponding to those employed in circuit investigations. With negligible dissipation due to losses in the source, air, and sink, the damping of a particular mode is an indication of the absorption of energy at the boundary. From this damping coefficient the effective absorption coefficient of the boundary under the conditions determined by the mode may also be obtained. The simplest result is obtained when the wave front is parallel to the walls on which it is

incident. The same apparent absorption is obtained from the transient decay of the same mode.

Likewise, by analogy with the electrical case, we may think of the transient current which flows through the microphone load when a unit d-c potential is suddenly supplied to the speaker terminals as the "transfer indicial admittance" of this electroacousto-electrical network. Viewed in this light the transient response of the speaker itself (a small part of the dynamical system) or the transfer indicial admittance of the loud-speaker and microphone mounted in a free field where no reflections are present takes on much less significance.

We know experimentally that any room which is considered a good acoustical listening environment has appreciable reverberation and therefore that the rate of decay of the energy in the resonators of the room is only moderate. Experimentally it has been found that the rate of decay of the modes of the speaker itself measured in a free field is of the same order. If the loud-speaker is loosely coupled to the room, *i.e.*, if the room impedance seen by its diaphragm is small as compared with the impedance seen looking into the diaphragm, then we may loosely think of the loud-speaker as converting the unit d-c emf into a number of damped sinusoidal terms (one corresponding to each mode of the speaker), which in turn excite the room. The spectral composition of a single damped sinusoid for two rates of decay is shown in Fig. 7. From these we see that, if the rate of decay is large, the number of room modes excited may be large, because of the broad frequency spread of the energy exciting the room. Actually, of course, we should only think of the dynamical system as a whole and the above explanation as a simplification of the problem.

Experimental curves of this transfer indicial admittance are what one would predict from the theory. If the speaker is loosely coupled to the room, if its damping when it looks into a small acoustical impedance is low, and if an undamped resonant frequency of the speaker lies near one of the resonant modes of the room, the transient term looks like a typical one for two coupled circuits. That is, there are two prominent rates of decay containing the combined effect of the two important resonant frequencies (the speaker and room). On the other hand, if the speaker is highly damped when looking into a high acoustical impedance, if the driving point impedance of the room at an undamped natural frequency of the diaphragm is high, and if the room is large or its dimensions are so chosen that there are a number of resonant frequencies near an undamped resonant frequency of the speaker diaphragm, then the transient term consists of the superposition of a large number of damped sinusoids. In this case it may be seen that elimination of the term due to the loud-speaker would cause a negligibly small change in the apparent transient response of the system. This was verified in an unreported investigation in which it was shown experimentally that, if the fundamental speaker mode was eliminated by the use of a properly chosen electrical network, the aural result in reproduced speech and music was small unless the damping of the speaker radiating into a free field was unusually small.

The more important practical implications of the above (see also Sec. 7, Mutual Radiation Impedance) are the following: (1) The loud-speaker should preferably be mounted in the corner of the room. In this position the greatest number of room resonances are "excited" and the most energy is supplied to the room. (2) The *average* l-f radiation is a maximum when the speaker is as near the floor (or ceiling) as possible and in the room corner. Next most desirable location is near floor (or ceiling) and side wall. (3) At any one frequency, maximum radiation is obtained when the room impedance seen by the diaphragm is the conjugate of the impedance seen looking back into the diaphragm, that is, for the location which makes the *combined* speaker and room resonate. Such a maximum may not be obtained if the longest room dimension is less than roughly a half wavelength long. (4) Because of this and the small number of resonant frequencies which occur in small rooms at low frequencies, small rooms do not normally permit the best l-f reproduction. (5) Corner positions also permit improved h-f response because of the smaller solid angle the radiation has to cover. (6) The l-f transient response of the speaker itself is

not so important as is generally supposed because the *desired* transient response of the room, which is a necessary property of a listening room judged to be good by trained listeners, helps obscure this "distortion."

OBJECTIVE LOUD-SPEAKER TESTS

The following more important characteristics of a loud-speaker must be determined in any complete test: response-frequency, efficiency-frequency, directional, impedance, and distortion.

33. Response-frequency Characteristic (Steady State). A response-frequency curve of a speaker is a curve graphically depicting the sound pressure produced at a designated position in the medium, the electrical input and acoustic environment being specified. Frequency discrimination is the most important form of distortion in many loud-speakers, and the response curve attempts to indicate quantitatively the amount present. Since the ear is primarily responsive to the sound pressure, the ordinate of the curve is made proportional to it or to its average value in a specified region.

The response curve is obtained by connecting the loud-speaker to a variable frequency source of specified internal impedance and constant specified internal voltage. The pressure at one or more points in the medium is measured as the frequency is varied slowly enough so the resulting measurement does not differ appreciably from the steady-state value.

A "free-field" response curve is made outdoors or in an anechoic ("dead") room in the absence of unintended reflecting surfaces and is probably the most useful single curve showing the loud-speaker performance for outdoor applications. Curves of this type are valuable because (1) the direct incident sound from the source in various directions may be accurately determined and a close estimate made of the direct sound indoors; (2) the acoustic environment is relatively simple since only intended reflecting surfaces are included. The efficiency of the unit may then be accurately obtained from this environment at some frequencies and estimated at others, since the impedance seen by the diaphragm will change slowly with frequency, *i.e.*, the transfer admittance of the speaker, air, and microphone is a smooth curve. (3) The specified test conditions may be duplicated relatively easily at various laboratories, permitting significant comparison of test results. The construction of identical normal listening or live test rooms, however desirable, would be difficult, partly because of differences of opinion on an "average" room and partly for economic reasons.

By "intended" reflecting surfaces is meant those that are an intended part of the radiating system. Frequently a cabinet or enclosure is measured outdoors in the absence of all reflecting surfaces, *i.e.*, radiating into a solid angle of 4π steradians or a complete sphere. This is usually undesirable since most enclosures are intended to operate on a floor and against a wall, which adds two intended reflecting surfaces and primary images (see Radiation Impedance). If the enclosure is intended to operate in a room corner, there are three important primary images. The impedance seen looking into these reduced solid angles of π and $\pi/2$ steradians, respectively, is very different from the 4π case and usually results in an error of the order of 6 to 8 db at low frequencies. The error is especially large in the case of compound sources such as vented enclosures. The measured nonlinear distortion usually differs by a much larger factor.

Outdoor measurements into solid angles of π (wall and floor) and $\pi/2$ steradians (corner) are made by constructing large rigid nonabsorbing surfaces.

Other intended parts of the speaker such as the baffle, horn, and enclosure should, of course, be specified. The normal impedance or the impedance looking into the signal terminals of the speaker with the acoustical load (acoustical environment), used when the response curve was obtained, should be plotted. Both the angle and modulus of this impedance are required if the response of the speaker with any source impedance other than that employed in the test is to be calculated.

If a space average of the pressure is obtained by moving the microphone or by using multiple microphones, details of the method should be given. If a warble tone or noise generator is used to get a "moving frequency average" of the transfer admittance, the spectral composition of the source should be specified. These expedients and the one involving motion of the loud-speaker, which is usually unsatisfactory, are recommended only for indoor measurements when the room does not provide approximately free-field conditions.

Normal listening-room measurements are made with the loud-speaker mounted in its intended position in a typical listening room. As noted under Room Acoustics, the room impedance seen by the loud-speaker depends on the characteristics and location of the source itself (diaphragm sizes, locations, and modes of vibration), the

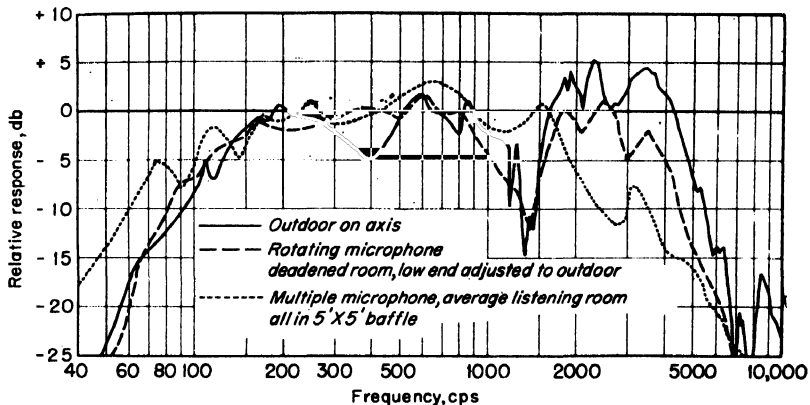



FIG. 48. Response of one speaker as measured by three different companies, showing the futility of relying on response curves without a complete knowledge of the specific test, acoustic environment, and method.

geometry of the room, and the impedance of its boundaries. This means that the energy supplied the room depends on the particular room and speaker location chosen. This is frequently used as an argument against this type of test. Since the results obtained in reasonably similar rooms, with similar speaker locations in each, differ by only a moderate amount, this disadvantage does not outweigh the many important advantages of this type of test, some of which are: (1) the impedance seen by the loud-speaker (including cabinet or enclosure), averaged over a small frequency interval, is closer to the average impedance seen under operating conditions than the impedance seen under the usual (4π steradians) outdoor test conditions; (2) calculation of the indoor from the outdoor performance is only of academic interest when hundreds of response curves are to be obtained, because of the labor involved; (3) ready comparison of the results of objective and subjective or listening tests in the same room is possible if the room is a good listening room; (4) one is not at the mercy of the weather; and, most important, (5) test facilities are readily provided in almost any organization.

Three large laboratories measured one speaker and plotted what they would publish as the response-frequency graph of the speaker. The results are shown in Fig. 48. This does not indicate any error in measurement. Actually different things were measured in each case. The curves indicate that response curves must be interpreted with great care and then only by a person familiar with the many factors involved. No experienced speaker designer or user thinks of choosing a speaker solely or even largely on the basis of a response curve.

34. Efficiency-frequency Characteristic. If the free-field-pressure response at a sufficient number of points on a spherical surface centered on the diaphragm is obtained, the total acoustical output may be calculated.

The efficiency-frequency curve of a speaker corresponds to the response-frequency curve except that the ordinate indicates the efficiency (usually "rated" or system). In a typical listening environment and listener location the direct incident sound energy, which would be approximately indicated by the free-field response-frequency curve at the listener's location with respect to the speaker, is only a small fraction of the reflected sound energy. A curve which gives the pressure, averaged over the useful listening region, then indicates the probable pressure the listener will experience. If absorption at the room boundary is independent of frequency this will be proportional to the total energy emitted by the speaker. Efficiency-frequency or space-averaged response-frequency curves are therefore very useful in interpreting indoor operation. Outdoor response-frequency curves at various angles off the speaker

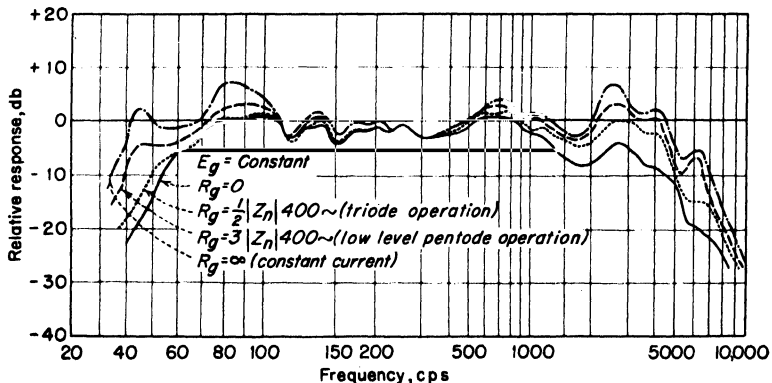



FIG. 49. Variation in total sound power radiated by a bass reflex enclosure into a room as source impedance (impedance match) is varied. Generator or source voltage constant for each curve but arbitrarily raised as the source resistance was raised to maintain constant acoustic output at 400 cycles.

axis, with the speaker radiating into approximately the solid angle it will see indoors, are also important since the listener weights the direct incident sound energy more heavily than the same energy if in a reflected wave. Haas[1] has shown that this arises from the fact that the initial wave takes control of the listening mechanism in such a way that a subsequent identical wave delayed by from 5 to 30 msec has to exceed the first by the order of 10 db, even if coming from a different direction, to be reliably identified as a separate disturbing signal. This also accounts for the fact that it is very desirable to so mount reinforcing speakers in speech systems that their air path distance to listeners is from 6 to 35 ft greater than the distance from the microphone to the listeners. The success of the directional speaker located some distance directly over or even a little behind the microphone in a voice-reinforcing system is accounted for primarily by this factor.

SUBJECTIVE LOUD-SPEAKER TESTS

35. Listening or Subjective Tests. Listening tests are a necessary part of the complete test of a loud-speaker. While physical measurements are valuable in design work and in quantitatively determining some performance criteria, they cannot at present completely specify the subjective performance. Because of the apparent simplicity of listening tests many important factors are frequently neglected, with the result that many tests are meaningless and others actually misleading.

In both indoor and outdoor tests all precautions should be taken that are used in

[1] Haas, v. H., Über den Einfluss eines Einfachechos auf die Hörsamkeit von Sprache, *Acustica*, **1** (2), 49, 1951.

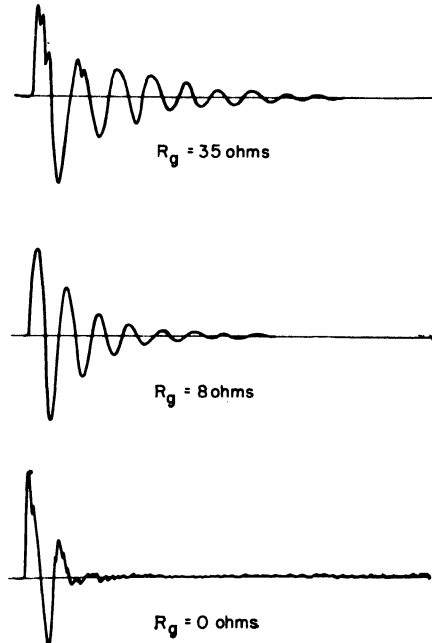
physical or objective tests. The only essential difference is that the listener is substituted for the microphone. The properties of the ear and listener must therefore be considered in interpreting the results.

36. Relative-loudness Efficiency. The most common test is one to determine the relative-loudness efficiency of two speakers. An attenuator in the amplifier which does not alter its response is adjusted (usually with a relay which also switches the speakers) to attenuate the input to the louder speaker by the amount required to make the speakers equally loud. The required attenuation of the louder in decibels is their relative loudness efficiency in decibels. The relative loudness will depend primarily on the spectral composition of the test signal, the response-frequency characteristic of the speakers, and on the sound intensity. Tests on the speaking and singing voice and various types of music are usually averaged. A valuable signal source for this and response-frequency tests is a "flat" noise source, or one in which the energy is uniformly distributed with frequency. This particular spectral composition ensures energy at all frequencies at which significant differences in the speaker response may occur.

FIG. 50. Response of moving-coil speaker with 8-ohm voice coil to an electrical pulse as source resistance is varied. Minor irregularities in curves are due to high-frequency modes of vibration of diaphragm.

37. Response-frequency Characteristic. Apparent subjective response-frequency tests may be made with the same signal source used in the loudness tests. Since the listener is not mobile, "space-averaging" methods employed with microphones cannot be used and "frequency-averaging" methods are employed. While noise sources are occasionally used in objective tests, they have unfortunately been neglected in subjective tests, where they are of special value because the trained ear can quickly appraise response differences which are missed if the signal source contains no energy at the frequencies at which differences occur.

38. Distortion Characteristic. Except with a single- or double-frequency input (the latter to determine intermodulation) it is difficult to determine the distortion characteristic of the speaker itself. With one or two simultaneously applied frequencies the input to the speaker is readily determined when the normal impedance of the speaker is known. This is not true of a signal of random energy distribution, and therefore with such a signal the apparent input to the speaker is not readily determined unless the normal impedance is relatively independent of frequency. When the *speaker* distortion characteristic is desired, the amplifier should be capable of supplying many times the rated input power to the speaker without distortion because of the high ratio of peak to average energy in speech and music (see these sections). Much overload charged to speakers is amplifier overload.

What is usually measured is the *system* distortion characteristic. Since amplifier overload almost invariably occurs at about the level at which speaker overload occurs in an economically planned system, what is evaluated is the combined system overload. In this case the speaker with the most restricted h-f response (other factors

being equal) will have the best *system* overload rating since annoying h-f distortion products are attenuated.

There is no standard for speaker input power rating, but in practice a speaker rated at X watts will "handle" the output of an X-watt amplifier, which is not overloaded, with speech or music (complex wave) input (see Motors and Secs. 1 and 18). The rating does not indicate the power the speaker will handle at a single frequency but takes advantage of the *normal* (no heavy bass or treble accentuation) spectral composition of speech and music. To avoid the trouble in determining the signal level across the speaker, with its variation in normal impedance, the grid voltage on the output stage may be measured. The signal input is raised until perceptible distortion results. The peak grid voltage is then measured with an indicator having a negligible time constant, such as a cathode-ray tube. A resistance equal in value to the magnitude of the nominal loud-speaker impedance is substituted for the speaker. The power dissipated in the resistance load with a 400-cps signal having the same maximum value as the maximum signal is considered (since it includes the output stage) the speaker input power rating.

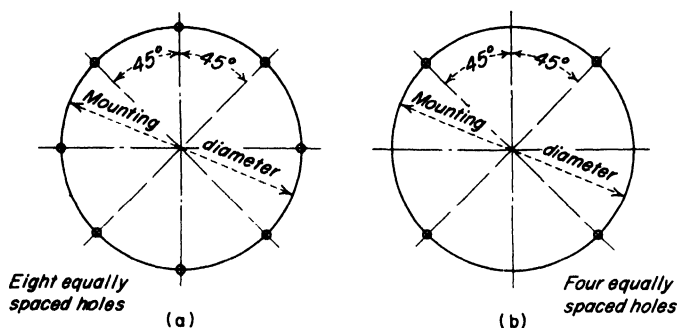



Table 4. EIA Standard Mounting Dimensions for Direct-radiator Speakers

Nominal speaker size, in.	Hole arrangement	Mounting diam, in.	Min hole diam, in.	Baffle* hole diam, in.
3½	*b*	3$^{15}/_{16}$	$^{3}/_{16}$	3⅛
4	*b*	4$^{11}/_{16}$	⅛	3¾
5	*a*	4$^{11}/_{16}$	⅛	4⅛
5¾	*a*	5⅜	⅛	4¾
6½	*a*	6⅛	⅛	5⅜
8	*a*	7⅝	$^{7}/_{32}$	6¾
10	*b*	9⅝	$^{7}/_{32}$	8¾
12	*b*	11$^{9}/_{16}$	¼	10¾
15	*a*	14$^{9}/_{16}$	$^{17}/_{64}$	13¼

* Not an adopted standard.

Intermodulation tests performed by measuring the distortion products generated when two sinusoidal signals of different frequency are applied are of value in design work but of limited value to the user because the results may be confusing or even misleading. If a high-amplitude fixed low-frequency signal (usually between 60 and 400 cps) is applied and a second signal, normally of constant but reduced voltage and variable from 1 to 20 kc, is also applied, through a low-distortion mixer, the intermodulation products can be measured. The converse may also be done; namely, apply a fixed h-f and vary the l-f signal. In a typical speaker these distortion products will be found to vary markedly with the frequency of the applied signals. Because of the marked frequency dependence preselected frequencies of the type

used in amplifier measurements may be very misleading. No single figure of merit based on the physical measurements has been found which correlates at all satisfactorily with the objectionability of the distortion products assessed by a listening jury.

39. Outdoor Tests. These should be conducted so the listener subtends the same or a known angle with each source. The sources should be mounted so their mutual-radiation impedance does not influence the result. Separating the sources by several times the diameter of the cone or horn mouth usually suffices. The energy absorbed by the unused speaker will be more nearly independent of frequency if the voice coil of the unused speaker is short-circuited. Unwanted reflecting surfaces should be avoided.

40. Indoor Tests. The speakers should be separated by several times the diameter of the cone or horn mouth to minimize mutual-radiation impedance. This is particularly true if the speakers are mounted on a common open baffle. Some coupling between the sources will always exist because of the transfer admittance between the two in the room (see Room Acoustics). It is important to mount the speakers symmetrically with respect to the room and listener in order to provide similar coupling between each source, the room, and the listener.

References

General:

American Standards Association, Z24.1-1942, Z24.2-1942, Z24.3-1944, Z24.4-1942, C16.4-1942.
Beranek, Leo L.: "Acoustic Measurements," Wiley, 1949.
———: "Acoustics," McGraw-Hill, 1954.
IRE Electroacoustic Standards, 1951.
J. Acoust. Soc. Am. (see Cum. Index).
J. Audio Engrg. Soc. (see Cum. Index).
McLachlan, N. W.: "Loudspeakers," Clarendon Press, 1934.
Morse, P. M.: "Vibration and Sound," McGraw-Hill, 1948.
Olson, H. F.: "Elements of Acoustical Engineering," Van Nostrand, 2d ed., 1947.
——— and F. Massa: "Applied Acoustics," Blakiston, 1939 ed.
Rayleigh, Lord J. W. S.: "Theory of Sound," Dover, 1945.
Soc. Mot. Picture Engrs., J.
Stewart, G. W., and R. B. Lindsay: "Acoustics."
Wien-Harms: "Handbuch der Experiment Physik," Vol. 17, Parts 1-3.

Speech, Music, and Hearing:

Beasely, W. C.: Characteristics and Distribution of Impaired Hearing in the Population of the United States, *J. Acoust. Soc. Am.*, **12**, 114-121, 1940; also National Health Survey, *Hearing Study Series Bulls.* 1-7.
Fletcher, H.: "Speech and Hearing"; also Physical Characteristics of Speech and Music, *Bell System Tech. J.*, **10**, 349, 1931; also Hearing, the Determining Factor in High Fidelity Transmission, *Proc. IRE*, **30**, 266, 1940; Steinberg, J. C.: Articulation Testing Methods, *Bell System Tech. J.*, **10**, 806, 1929; and Munsen, W. A.: *J. Acoust. Soc. Am.*, **5**, 83, 1933.
Sivian, L. J.: Speech Power and Its Measurement, *Bell System Tech. J.*, **10**, 646, 1939; and Dunn, H. K., and S. D. White: "Absolute Amplitudes and Spectra of Musical Instruments and Orchestras."
Snow, W. B.: Audible Frequency Ranges of Music, Speech and Noise, *J. Acoust. Soc. Am.*, **3**, 155 166, 1931.
Stevens, S. S., and H. Davis: "Hearing, Its Psychology and Physiology."

Loud-speakers:

Bostwick, L. G.: *Bell System Tech. J.*, **8**, 135, 1929. (Tests.)
Cook, E. D.: *Gen. Elec. Rev.*, **33**, 505, 1930. (Tests.)
Hall, W. M.: *J. Acoust. Soc. Am.*, **3**, 552, 1932. (Horns.)
Hunt, F. V.: "Electroacoustics," Wiley, 1954.
Klapman, S. J.: Interaction Impedance of a System of Circular Pistons, *J. Acoust. Soc. Am.*, **11**, 289, 1940.
Knowles, H. S.: *Electronics*, **4**, 154, 1932; also **6**, 240, 1933. (Tests.)
Olney, B.: *Proc. IRE*, **10**, 1113, 1931. (Tests.)
Pedersen, P. O.: *J. Acoust. Soc. Am.*, **6**, 227, 1935; and **7**, 64, 1935. (Subharmonic theory.)
Phelps, W. D.: *J. Acoust. Soc. Am.*, **12**, 68, 1940. (Horn losses.)
Rice, C. W., and E. W. Kellogg: *J. AIEE*, **44**, 982, 1015, 1925. (Moving-coil speaker.)
Salmon, V.: *J. Acoust. Soc. Am.*, **17**, 199, 212, 1946. (Horn theory and new horns.)
Stenzel, H.: *Elektrische Nach. Tech.*, **4**, 239, 1927; **6**, 165, 1929; and **7**, 90, 1930; also *Z. Tech. Physik*, **10**, 569, 1929; and *Ann. Physik*, **11**, 947, 1930. (Directivity of radiators.)
Wegel, R. L.: *J. AIEE*, **49**, 791, 1921. (Theory.)

Wolff, I., and L. Malter: *Phys. Rev.*, **33**, 1061, 1929. (Mutual radiation impedance.) Also *J. Acoust. Soc. Am.*, **2**, 201, 1930. (Directivity.)

Room Acoustics:

Acoustical Materials Association Bulletins.
J. Acoust. Soc. Am., Symposium, July, 1940, Bibliography (see also Cum. Index, **1–10**).
Knudsen, V. O.: "Architectural Acoustics," 1932.
Morse, P. M., and R. H. Bolt: Sound Waves in Rooms, *Revs. Modern Phys.*, **16**, 69, 1944.
Sabine, P. E.: "Acoustics and Architecture," 1932.
Watson, F. R.: "Acoustics of Buildings," 1930; and *J. Acoust. Soc. Am.*, **3**, 14–43, 1931.

CHAPTER 12

NONLINEAR CIRCUITS

By Keats A. Pullen[1]

1. Definition. A linear circuit has a transfer function (ratio of output to input) which is not a function of any characteristics of the input signal (amplitude, frequency, or phase, for example); it does not introduce any variations or alterations of the shape of the wave. It may introduce a time delay, that is, it may repeat the wave exactly but with a small delay. Conversely, a nonlinear circuit has a transfer function which is dependent on one or more characteristics of the input signal. Whereas the linear circuit introduces a constant (or zero) time delay and no change in the wave form, the nonlinear circuit may introduce a variable delay and a change of wave form.

2. Types of Nonlinear Circuits. The types of nonlinearities which can be obtained with conventional circuits may be divided into four principal groups:

1. Frequency
2. Phase
3. Amplitude
4. Time-induced amplitude, or LTV (linear time variant, Sec. 89)

In actual practice, it is often difficult to classify some nonlinear conditions exactly because two or more types of conditions may occur simultaneously. For that reason, the discussion considers separately each particular type of nonlinearity in relation to its prime applications. For example, most devices which generate frequency distortion also produce phase distortion. In a similar manner, a device capable of generating amplitude distortion may also produce an amplification which varies with time, thereby introducing a time-controlled type of amplitude nonlinearity.

3. Frequency Nonlinearities. Circuits having nonlinear frequency characteristics show an amplification which varies as a function of frequency. In many of the more routine applications of electronics, for example, amplifiers and related devices, a circuit is intentionally adjusted to have a minimum variation of amplification with frequency over a specified range (the amplification may be less than unity and represent a loss). There are also many applications in which the variation of amplification with frequency is used to accomplish specific results; it is the purpose of this and the following sections to show typical applications and briefly note the techniques used in designing circuits to provide a variable amplification as a function of frequency.

The simplest circuits having a frequency type of nonlinearity normally show a constant amplification over one range of frequency and an amplification which changes by a ratio of 2:1 for each octave change of frequency in another range (change of frequency to either half or twice the initial value).

Devices which possess one type of characteristic in one range of operation and another type of characteristic in another range have a transition region in which neither characteristic alone applies. In this region the characteristic is a combination of the two. The *critical frequency* or *transition frequency* is that frequency at which the impedances of the two elements are equal, or $|Y_1Z_2| = 1$. It separates the two ranges and specifies the intersection point of the tangents to the two sections of a plot, on a loglog scale (Fig. 1) of the amplification as a function of frequency. The transition region extends from just under half to just over twice the critical frequency.

[1] Aberdeen Proving Grounds.

4. Integrator Circuits. Circuits which offer no hindrance to the passage of signals having frequencies well below the critical frequency but which attenuate the amplitude to half for each doubling of the frequency above the critical frequency may be called integrating circuits. In their simplest form, there are four basic combinations

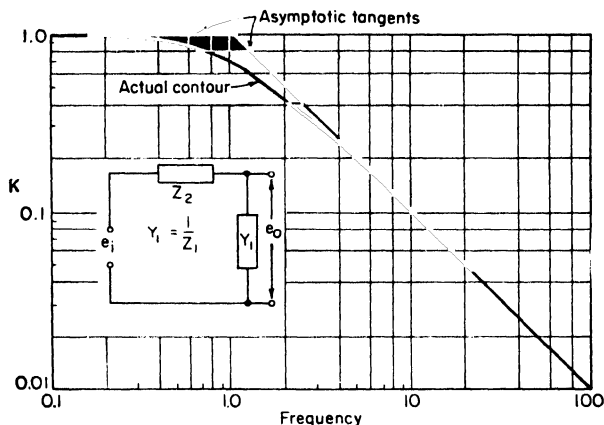



FIG. 1. Basic L-section network.

of R, C, L, or constant-voltage or -current sources, used in pairs, which may be used to form integrator circuits. They are:

1. RC integrators (passive)
2. LR integrators (passive)
3. Capacitive constant-current integrators (active)
4. Inductive constant-voltage integrators (active)

Each of these is discussed in turn after an examination of the basic operating conditions for the series-shunt, or L-section, circuit.

5. Series-shunt, or L-section, Circuit. The function of this type of circuit may be determined, as indicated in Fig. 1, in terms of the equation

$$K = 1/(1 + Y_1Z_2) \tag{1}$$

where K is the amplification, or $K = e_0/e_i$. In this equation, one of the parameters is resistive, the other either capacitive or inductive. Any combination which makes the Y_1Z_2 term proportional to frequency yields an integrating circuit; if this term varies inversely with frequency, a *differentiating circuit* results. The critical transitional frequency, as noted above, is given by

$$|Y_1Z_2| = 1 \tag{2}$$

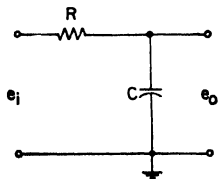



FIG. 2. Integrating circuit.

These equations are satisfied for an integrating circuit with Y_1 capacitive and Z_2 resistive, or Y_1 resistive and Z_2 inductive.

6. RC Integrators. The RC integrator (Fig. 2) uses a series resistor and a shunt capacitor to produce integration. Integration occurs under conditions for which Y_1Z_2 is large compared with unity, with the result that the critical frequency must be less than a fifth of the lowest frequency to be handled by the integrator. As can be seen, Eq. (1) takes the form $1/pCR$ when $pCR \gg 1$, where the $1/p$ indicates an integration. Physically, this integration is obtained by the charge stored in the capacitor

suppressing the voltage variations which are applied to the series resistor. The integrator tends to have a constant voltage output.

7. *LR* Integrators. The *LR* integrator uses a shunt resistor and a series inductor to produce integration. As with the *RC* integrator, integration occurs when the value of Y_1Z_2 is large compared with unity. Once again, the critical frequency should be less than one-fifth of the lowest frequency to be handled by the integrator. The equivalent integration equation for the *LR* integrator takes the form $1/pGL$, where $G = 1/R$. Physically the integration is a consequence of the induced voltage (derivative effect) developed in the inductor minimizing the rapid changes of current in the shunt resistor, with the result that a constant-current output is generated for frequencies above critical.

8. Capacitive Constant-current Integrators. Integration may be obtained by the use of passive elements in conjunction with active devices as well as with passive circuits alone. Here the active device serves as a constant-current source, thereby charging and discharging a capacitor (X_L) used as a load impedance, as in Fig. 3, and producing direct integration. The main advantage of the active integrator is the greater voltage efficiency available. A signal voltage is applied to the active device for conversion into a current source. Using a pentode tube as the source with the screen by-passed to its cathode and with cathode degeneration, the current equation takes the form, based on Fig. 3,

$$i_o = g_{m1}e_i/(1 + g_{m1}R_k) \tag{3}$$

where g_{m1} is the transconductance of the tube and the other factors are defined in Fig. 3. The output voltage takes the form

$$E_o = i_o/pC \tag{4}$$

where i_o is the output current. This circuit arrangement is one of the best simple integrators for many applications. More complex, and more accurate, integrators can be obtained through the use of high-gain amplifiers with integrator circuits in the feedback paths.

FIG. 3. Active integrating circuit.

9. Inductive Constant-voltage Integrator. This integrator uses a stabilized voltage amplifier in connection with a conventional *LR* integrator to develop an output current which is the integral of the applied voltage. Its operation is identical with that of the *LR* integrator already discussed.

10. Differentiating Circuits. Circuits which offer little or no hindrance to the passage of signals having frequencies well above the critical frequency but attenuate to half for each octave below this frequency may be called differentiating circuits. In their simplest form, there are four basic combinations of *R*, *C*, *L*, and constant-voltage or constant-current sources, used in pairs, which may be used to form differentiator circuits. They are:

1. *CR* differentiators (passive)
2. *RL* differentiators (passive)
3. Inductive constant-current differentiators (active)
4. Capacitive constant-voltage differentiators (active)

The equation governing the operation of the simple differentiator (using the series-shunt circuit) is exactly the same in form as for the integrator circuit, viz.,

$$K = 1/(1 + Y_1Z_2) \tag{1}$$

To obtain differentiation, however, the Y_1Z_2 term in the denominator is inversely proportional to frequency, or takes the form

$$Y_1Z_2 \sim 1/p \tag{5}$$

The critical, or transition, frequency occurs when the YZ product has a magnitude of unity, and differentiation is obtained when the value of the product is large compared with unity, a minimum of five or more. The amplification equation then takes the form

$$K = p/A \tag{6}$$

where A is a dimensionless quotient, either $1/RC$ or R/L.

11. *CR* Differentiators. These make use of a capacitor as the series element Z_2 and a resistor as the shunt element Y_1. This arrangement permits the transmission of sharp voltage changes to the resistor, but blocks the passage of slow changes. The value of A in Eq. (6) then is $1/RC$, making Eq. (6) read

$$K = pCR \tag{7}$$

12. *RL* Differentiators. The RL differentiator has the reactance element as the shunt element, using the build-up of impedance with frequency to cause the increased output. The amplification equation takes the form

$$K = pL/R \tag{8}$$

13. Inductive Constant-current Differentiators. Differentiation can also be obtained by use of a constant-current tube with a reactance in its plate circuit which increases linearly with frequency. Such a differentiator may be assembled by use of a pentode constant-current tube having an inductance for Z_L of Fig. 3. The gain is then

$$K = g_{m1}pL/(1 + g_{m1}R_k) \tag{9}$$

where Z_L of Fig. 3 has the reactance pL.

14. Capacitive Constant-voltage Differentiators. This differentiator actually consists of a voltage amplifier followed by a CR differentiator; as a result it need not be discussed further. Its operating equation, which consists of the products of those for the two components, may be written

$$K = K_0pCR \tag{10}$$

APPLICATIONS OF INTEGRATORS AND DIFFERENTIATORS

Integrating circuits are used extensively as filters to reduce coupling from power-supply impedance and to reduce the amount of ripple and noise which can get to the input circuits of high-gain amplifiers. The RC form is particularly effective because of its ability to supply the charge storage required to provide the variable currents used by electronic circuits.

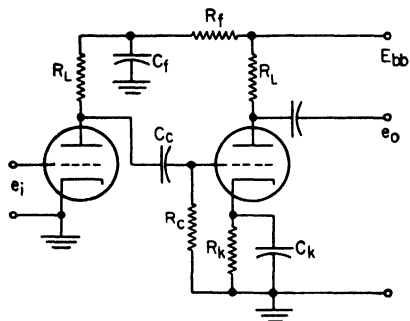



Fig. 4. Decoupled amplifier.

15. The Decoupling Filter. RC filters are commonly used between amplifier stages to minimize coupling through the impedance of the power supply, and to reduce ripple and noise. These filters can also prevent oscillation and motor-boating. A typical circuit is shown in Fig. 4.

The signal-frequency loss of the decoupling network for the high-voltage supply of an amplifier should be sufficient to make the product of the loss factor K_d in the filter and the forward gain of the amplifier K less than unity over the full range of operation if self-oscillation is to be prevented. In addition, the capacitance should be large enough to keep the voltage variation across it small compared with the output signal voltage (less than a fifth of

the output). These limiting conditions are

$$KK_d = K/\omega_1 C_f R_f < 1 \tag{11}$$

and

$$\omega_1 C_f R_L = 5 \tag{12}$$

where R_L is the load impedance and C_f and R_f are the filter components. The capacitor size C_f given in the equation is the minimum which normally can be used without introducing low-frequency distortion and phase shift. For applications requiring maximum constancy of delay, an even larger value of capacitance should be chosen, from two to five times that given by the equation.

A decoupling filter (C_sR_s in Fig. 5) may be used to limit voltage variation developed at the screen of the amplifier tube. In this application, the filter is used for two purposes, the first being to reduce the supply voltage E_{bb} and the second to keep the screen voltage sufficiently constant that screen degeneration is not introduced. The value of R_s used must be such as to reduce E_{bb} to the screen voltage E_{c2} at the average

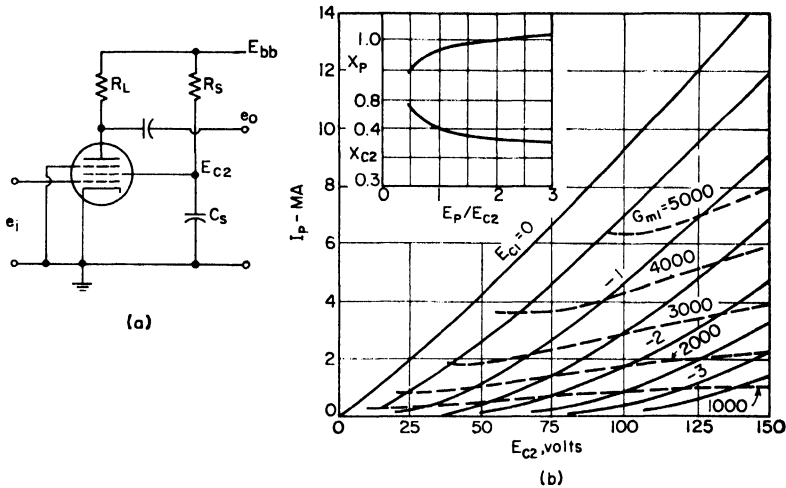



Fig. 5. Pentode amplifier.

screen current drawn by the tube; the value of C_s should be large enough to keep the variation of the screen voltage under full signal conditions from exceeding the input voltage. (This is a rule of thumb which works out effectively.) The exact equation may be found, in terms of screen-to-plate transconductance, to reduce to the approximate form

$$G_{m2}X_{c2}/\omega_1 C_s = 0.2 \tag{13}$$

where the 0.2 has been chosen to make certain that degeneration will not alter the over-all amplification by more than a few per cent. In this equation, X_{c2}, the screen correction factor which gives the percentage of the nominal plate current drawn by the screen, has a value between 0.1 and 0.4 for most pentodes, G_{m2} is the slope of the screen voltage–plate current curve,[1] ω_1 is the radian frequency at which the low-frequency response of the amplifier is reduced by 3 db, and C_s is the screen by-pass capacitance.

The equation for the value of screen resistance required may be written as

$$R_s = (E_{bb} - E_{c2})/I_pX_{c2} \tag{14}$$

[1] Pullen, K. A., Conductance Curves Speed Pentode *RC* Amplifier Design, *Tele-Tech*, July, 1953; The Use of Screen-to-plate Transconductance in Multigrid Tube Circuit Design, *Elec. Eng.*, October, 1954.

where I_p is the average plate current, E_{c2} is the screen voltage, and X_{c2} is the screen correction factor.

The technique used for obtaining the operating screen and plate currents using the correction factors X_p and X_{c2}, sometimes called the conductance technique because it also facilitates the determination of transconductance values, is designed to give emphasis to the more important variations in the tube and to reduce to secondary status the less important relations. With pentodes, the important relations are the variation of the space current with screen voltage at fixed values of bias and the variation of the space transconductance with bias and screen voltage. These relations are presented on conductance data sheets in terms of a nominal plate current and a nominal transconductance to the plate (nominal because the value given is that which is obtained with the plate voltage twice the applied screen voltage); the nominal values are identified by the symbols I_p and G_{m1} or G_{m2}. The actual plate and screen currents and the transconductances are related to the nominal values by the following equations:

$$i_b = I_pX_p \qquad g_{m1} = G_{m1}X_p \qquad g_{m2} = G_{m2}X_p$$
$$i_{c2} = I_pX_{c2} \qquad g_{m12} = G_{m1}X_{c2} \qquad g_{m22} = G_{m2}X_{c2}$$

where i_b = instantaneous plate current
i_{c2} = instantaneous screen current
I_p = nominal plate current
g_{m1} = actual transconductance from first grid to plate
g_{m12} = actual transconductance from first grid to screen
g_{m2} = actual transconductance from screen to plate
g_{m22} = actual screen conductance
G_{m1} = nominal transconductance from first grid to plate
G_{m2} = nominal transconductance from screen to plate
X_p = plate correction factor for a given ratio of plate to screen voltage
X_{c2} = screen correction factor for a given ratio of plate to screen voltage

These definitions have been patterned after the standard IRE definitions and have been adapted as required to provide efficient organization and use. The modification of the symbols for static and limit values within the style of the IRE standard is explained in the footnote to Sec. 47.

One point to be remembered in the design of screen decoupling circuits for r-f and i-f amplifiers is that sometimes, particularly where AVC is used, the frequency ω_1 may be in the audio region instead of being about half the r-f or i-f frequency. This is because the presence of either hum or AVC in the power-supply voltage can leak into the screen circuit and, if G_{m2} changes sufficiently rapidly with E_{c2}, can actually modulate the signal being amplified, thereby establishing an oscillation.

A few other applications of simple RC integrators are listed below:

1. AVC decoupling circuits
2. Filters for use with detectors
3. Power-supply filters

16. Integrators with F-m Transmitters. If a phase-modulation transmitter is modulated with the integral of the desired communication information, the over-all result will be the production of a frequency-modulated wave. The frequency stability which can be obtained with such a transmitter greatly surpasses that of ordinary f-m type units. The integrator must be designed to operate efficiently and reasonably accurately over a frequency range of at least 1,000 to 1, with the result that the RC integrator must have carefully selected components. The capacitor must have a high Q and a small leakage over the frequency range, and the resistor must be stable and have a low shunt capacitance.

A preemphasis circuit is often used to permit an improvement of the over-all signal-to-noise ratio of the system. Possible ways of introducing the required preemphasis are the use of a differentiator to offset the integration, a by-pass circuit which makes the integrator partially ineffective above a selected high audio frequency, and the use of a zero-phase-shift circuit which has negligible loss below the audio range, a rapidly

increasing loss with frequency to the selected high frequency, and a constant loss above this frequency. The zero-phase-shift circuit is described in Sec. 26.

17. Uses in Wave Shaping. Integrator and differentiator circuits are very effective for wave shaping, since they can be used to suppress either the high- or the low-frequency components of an applied signal as is required. The integrator tends to emphasize the sine wave corresponding to the lowest-frequency component in the signal, converting a square wave first into a triangular wave, a triangular wave into a parabolic wave, etc., as shown in Fig. 6. The differentiator, on the other hand, emphasizes the higher-frequency components at the expense of the lower, converting a square wave into a pulse approximating a delta function, and on repeated differentiations creating waves somewhat as shown in Fig. 7. The number of zero-axis crossovers from each step of the square wave with a string of differentiators is one less than the number of differentiator circuits.

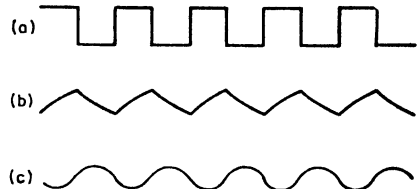



FIG. 6. Effects of integration on wave form.

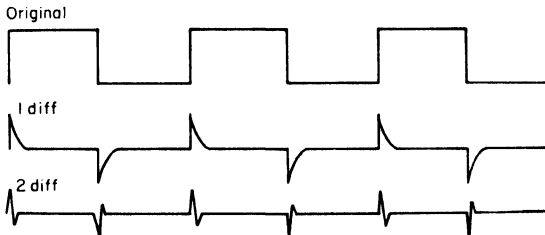



FIG. 7. Effects of differentiation.

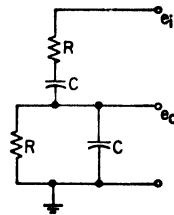



FIG. 8. *RC* frequency-selection circuit.

18. Integrators vs. Differentiators. In practice, integration usually can be accomplished with more nearly ideal results than differentiation for the following reasons:

1. The critical frequency usually can be placed well below the lowest required frequency with the integrator; it cannot always be placed above the highest with the differentiator because of the absence of an upper limit in the Fourier spectrum of an ideal wave form.
2. A high-frequency spectrum limitation on the incoming signal with a differentiator is as destructive to wave shaping as too low a critical frequency would be. A corresponding condition seldom occurs with integrator circuits.
3. Circuits for the improvement of integration linearity are simpler to design and build than are the corresponding circuits for differentiation.

Both integrator and differentiator circuits are essentially frequency-selection circuits. Integrators, when used as power-supply or decoupling filters, are intended to pass a narrow band of frequencies adjacent to zero frequency. Differentiator circuits are used as high-pass filters in the coupling networks in amplifiers. Consequently, the design of the corresponding circuits has to be coordinated if the best possible performance of the amplifier is desired.

Combinations of differentiator and integrator circuits can be used either for rejection or for emphasis of narrow bands of frequencies. Band rejection is accomplished by paralleling an integrator and a differentiator; the critical frequency for the differentiator should be appreciably higher than that for the integrator. Band pass may be obtained by having the critical frequency of the integrator higher than that of the differentiator and connecting the networks in series. A circuit showing a slight modification of such an arrangement as used in an oscillator is shown in Fig. 8. Most

of the Wien bridge type of *RC* oscillators make use of a combination which behaves as an integrator and a differentiator in series, the critical frequencies of each being chosen to give the required frequency characteristics.

RLC CIRCUITS

The *RLC* circuit in one of its many combinations is probably the most flexible of all frequency nonlinear or frequency-selective circuits. An *LC* circuit, if it could be made of lossless components, would be an optimum device for taking either the second integral or the second derivative of an applied wave form; since, however, the loss in the inductance is sufficient to make the circuit nonideal in behavior, the *RLC* circuit is used primarily for its frequency-selection properties and as a filter for power supplies.

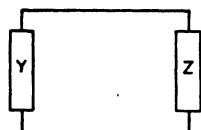



FIG. 9. Resonant circuit consisting of dual elements.

19. Phenomena in *RLC* Circuits. The operating characteristics of an *RLC* circuit by itself are determined by shock or transient excitation. If two low-loss circuits (or circuits having a dissipation per cycle small compared with the stored energy) which are duals of one another are interconnected with energy stored in one of them, as in Fig. 9, a transient vibration, damped in nature, commences and continues until all the stored energy is dissipated. (Two networks are duals if the susceptance of one is proportional to the reactance of the other. Resistive and conductive components need not be considered.) The energy in the one circuit is transferred to and stored in the other and then returned in a cyclical manner. In the strictest sense of the word, the true resonant frequency of the system is controlled by this free and controlled energy interchange. If the system is driven the response of the network depends on both the component arrangement and the way in which the circuit is excited.

FIG. 10. Series *LC* circuit.

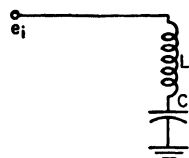
20. Driven *RLC* Circuits. Two principal forms of driven *RLC* circuits are commonly used—in fact, most frequency-selection circuits for *LC*-type oscillators can be represented as the one or the other. One of these is the series- and the other is the parallel-resonant circuit. Each of these circuits may be excited in a variety of ways, and each way causes the circuit to appear to have its own special frequency of maximum response. An example in terms of the series-resonant circuit is given below.

The circuit of Fig. 10 is used primarily as a trap in most electronic applications other than oscillators. It has the disadvantage of having a low impedance at resonance, with the result that it gives a low voltage output with ordinary conductance-controlled devices such as tubes and transistors. Its electrical characteristics make it particularly useful with voltage output, current input, or resistance-controlled devices.

FIG. 11. Typical plot of Z vs. ω for series-resonant circuit.

The general shape of the impedance curve for a simple series-resonant circuit takes the form shown in Fig. 11. Its equation is

$$Z = \sqrt{R^2 + (\omega L - 1/\omega C)^2} \quad (15)$$

Under constant-voltage excitation with the current $I = E/Z$, a maximum current will flow when the net reactance is zero.

Although the curves for Z and I are symmetrical when plotted on a logarithmic

frequency scale, the voltages in the circuits as actually used seldom show a corresponding symmetry. The voltage is developed by the current flowing through either an inductive or a capacitive reactance. Because of the symmetry of the current and the asymmetry of the reactance, the voltage is asymmetrical, as can be seen in Fig. 12.

FIG. 12. Typical variation of X_L, I_L, and E_L with ω.

The parallel-resonant circuit (Fig. 13*a*) is ideal for use with tubes and transistors operating under normal conditions because of their current-output characteristics. Parallel circuits may be excited either by a constant current or through magnetic coupling (which in effect couples a voltage source in series with the coil, giving the equivalent of constant-current excitation), or through a tap on the coil or capacitor as in Fig. 13*b* or *c*.

FIG. 13. Typical coupling circuits.

If the parallel-resonant circuit is excited from a constant-voltage source, either magnetic coupling or an inductive or capacitive tap is required to get the circuit excited in series. Then the output may be taken in parallel across either the coil or the capacitor as indicated.

21. Wave Filters. In addition to the normal applications of *LC* elements in tuned circuits, they may also be used in wave filters. In simple filters (constant-*K* type), the series element *Z* and the shunt element *Y* are dual elements or dual combinations of elements. When terminated in the proper resistances, such filters pass signals over some ranges of frequencies and reject them over others. Figure 14 and Table 1 indicate the characteristics and the controlling equations of the commonest types. Low-pass filters of the constant-*K* type are often used for power-supply filters.

FIG. 14. Constant-*K* wave filters.

The value of the terminating impedance as a function of frequency that is required by the simple wave filter often is not sufficiently constant to satisfy many applications. Consequently, a filter section called the *M*-derived section is commonly used (Fig. 15).

Table 1. Constant-K Wave Filters

Type	Z_1	Y_2	ω_1	ω_2	ω_0	ω_C	Z_T
Low.......	ωL	ωC	0	$\frac{2}{\sqrt{LC}}$	....	......	$\sqrt{\frac{L}{C}}$
High.......	$\frac{1}{\omega C}$	$\frac{1}{\omega L}$	$\frac{1}{2\sqrt{LC}}$	∞	....	......	$\sqrt{\frac{L}{C}}$
Band pass..	$\omega L - \frac{1}{\omega C}$	$K\left(\omega C - \frac{1}{\omega L}\right)$	$\frac{1}{\sqrt{KLC}}(\sqrt{1+K} - 1)$	$\frac{1}{\sqrt{KLC}}(\sqrt{1+K} + 1)$	$\frac{1}{\sqrt{LC}}$	$\frac{2}{\sqrt{KLC}}$	$\sqrt{\frac{L}{KC}}$
Band rejection	$\frac{\omega L}{\omega^2 LC - 1}$	$K\left(\frac{\omega C}{\omega^2 LC - 1}\right)$	$\frac{1}{\sqrt{KLC}}(\sqrt{1+K} - 1)$	$\frac{1}{\sqrt{KLC}}(\sqrt{1+K} + 1)$	$\frac{1}{\sqrt{LC}}$	$\frac{2}{\sqrt{KLC}}$	$\sqrt{\frac{L}{KC}}$

ω_1 = angular frequency for cutoff at lower limit
ω_2 = angular frequency for cutoff at upper limit
ω_0 = mid-angular frequency for band-pass filters
$\omega_C = \omega_2 - \omega_1$
Z_T = theoretical terminating impedance
K = arbitrary constant which affects band width and impedance

These sections are also used to increase the rejection at frequencies near the band edge where sharp cutoff is required. The basic equations for low- and high-pass filters of the M-derived type are included in Table 2.

As can be seen from the two tables, wave filters may be used for three principal applications, namely low-pass, high-pass, and band-pass. Only rarely are band-rejection filters required. Low-pass filters may be used as delay lines for video pulses, and band-pass filters for r-f carrier pulses. The delay characteristics (phase vs. frequency) for such filters are discussed briefly below under phase nonlinearities.

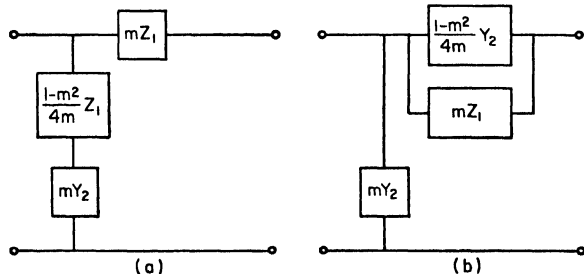



FIG. 15. M-derived half sections.

22. Tuned Circuits for Amplifiers. The design of tuned circuits for r-f and i-f applications is simple, but generally rather poorly understood. Two variables are available for adjusting these circuits, the reactance and the Q. The Q controls the band width, and the product of the reactance and the Q controls the impedance level, amplification K, and the stability. The equations are

$$Z = X_c Q \tag{16}$$
$$K = -g_{m1} Z \tag{17}$$

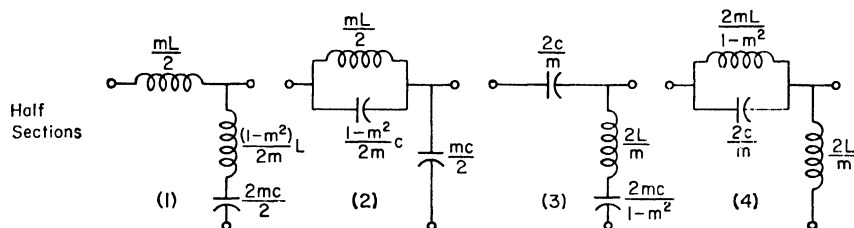
where Z is the resonant impedance of the tuned circuit. The value of K should be chosen sufficiently small that the product of K by the feedback gain K_f is less than unity (Fig. 16). The value of K_f should be less than

$$K_f = 1/\omega C_f Z \tag{18}$$

where C_f is the feedback capacitance as shown in Fig. 16. Equations (17) and (18) give a limiting value for Z in the form

$$Z = \sqrt{1/\omega C_f g_{m1}} \tag{19}$$

Table 2. M-derived Low- and High-pass Filters

Type	Z_1	Y_2	ω_1	ω_2	ω_∞	Z_m
1. Low.......	$m\omega L$	$\frac{4m\omega C}{\omega^2(1 - m^2)LC - 4}$	0	$\frac{2}{\sqrt{LC}}$	$\frac{\omega_2}{\sqrt{1 - m^2}}$	$\sqrt{\frac{L}{C}}$
2. Low.......	$\frac{4m\omega L}{\omega^2(1 - m^2)LC - 4}$	$m\omega C$	0	$\frac{2}{\sqrt{LC}}$	$\frac{\omega_2}{\sqrt{1 - m^2}}$	$\sqrt{\frac{L}{C}}$
3. High.......	$\frac{m}{\omega C}$	$\frac{4m\omega C}{4\omega^2LC - (1 - m^2)}$	$\frac{1}{2\sqrt{LC}}$	∞	$\omega_1\sqrt{1 - m^2}$	$\sqrt{\frac{L}{C}}$
4. High.......	$\frac{4m\omega L}{4\omega^2LC - (1 - m^2)}$	$\frac{m}{\omega L}$	$\frac{1}{2\sqrt{LC}}$	∞	$\omega_1\sqrt{1 - m^2}$	$\sqrt{\frac{L}{C}}$

ω_∞ = angular frequency for maximum attenuation
m = parameter which controls constancy of terminating impedance and frequency of maximum attenuation. A value of 0.6 gives best uniformity of the filter terminating impedance
Z_m = theoretical terminating impedance of filter

Actually, if the amplifier has only one or two stages, Z may sometimes be somewhat greater than the value given by Eq. (19) since the effect of phase has been neglected entirely. In multistage amplifiers, however, phase shifts from random causes can give with comparative ease the required phase for oscillation.

23. *RLC* Circuits as Coupling Devices. Inductive coupling circuits for coupling to or from coaxial lines may be either tuned or untuned (Fig. 17). Either series- or parallel-tuned circuits may be used; usually the Q of the coupling circuit is made low to broad-band the coupling circuit. Since the adjustment of any of these types of coupling loops may be accomplished with little difficulty if proper precautions are taken, the next few paragraphs indicate how advantage can be taken of the broad optimum which is available.

FIG. 16. Tuned amplifier with capacitive feedback.

As tuned circuits are coupled to one another, more and more of the energy in the excited circuit is withdrawn through magnetic coupling by the coupled circuit, reducing effective Q of each circuit. As a result, when critical coupling is reached, the optimum voltage coupling is reached, and the tuned circuits "double-hump" as the coupling is tightened further. This double-humping effect results in a limitation of the transformation ratio by developing an area between the two peaks in which the ratio is somewhat less than maximum. The two humps bracket the original operating frequency. Tightening the coupling toward unity moves one of the peaks toward zero and the other toward infinite frequency, giving an area of relatively uniform response in the middle.

The effective use of a tuned circuit as a coupling element requires the design to be such that the internal loss is small, and at the same time broad-band operation makes necessary a high over-all loss from the load. Consequently, the coupling circuit should have an unloaded Q in excess of 50, and a loaded Q between 2 and 10. The equations for use in the design of series-tuned coupling loops under these conditions are

$$\begin{array}{lll} Q_L = X_c/R_t & 2 < Q_L < 10 & Q_U > 50 \\ \omega^2 LC = 1 & K = \sqrt{Q_L Q_T} & \end{array} \qquad (20)$$

where R_t is the impedance of the load connected to the coupling circuit, X_c is the capacitive reactance of the series capacitor, Q_L and Q_U are the loaded and unloaded Q values for the loop, respectively, Q_T is the Q value for the tuned circuit, and K is the coefficient of coupling. For the shunt-tuned coupling circuit, only the equation for the loaded Q value is changed:

$$Q_L = R_t/X_c \qquad (21)$$

After the proper L and C values have been selected and the circuit assembled, the final adjustment is to tighten the coupling between the main tuned circuit and the coupling circuit until the desired operating characteristics are obtained.

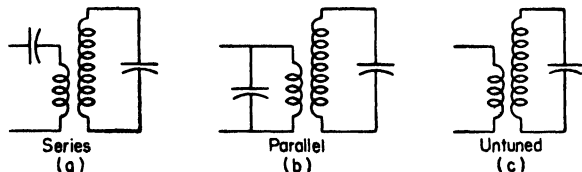



FIG. 17. Inductive coupling circuits.

A plot of the voltage-transfer characteristic as a function of the reactance of the untuned loop shows that a broad maximum occurs when its reactance X_L is approximately equal to R_t. This maximum is sufficiently broad that extensive charts can be prepared to indicate the relation, but frequently the charts are not needed if the maximum condition is used for design purposes. (The existence of the maximum may be shown mathematically.)

If the coupling between an untuned loop and a tuned circuit is gradually tightened, characteristics much similar to those of the tuned loop are noted. First the voltage transfer rises, finally levels, and then changes but little. The concept of critical coupling applies approximately with the untuned loop since Q_L now is taken as unity. (It is designed so that $X_L = R_t$.) The coupling equation becomes

$$K = \sqrt{Q_T} \qquad (22)$$

Physically, the process is as follows: first, as the coupling is tightened, transfer rises, but the loading from the coupled circuit is small compared with the damping of the tuned circuit itself. As the coupling increases, however, the damping due to the coupling link finally becomes equal to the damping in the tuned circuit; then the Q of the tuned circuit drops as coupling is increased, keeping the over-all transfer constant.

PHASE NONLINEAR CIRCUITS

The phase characteristics of signals are of rapidly growing concern, partly because of their importance in television, and particularly in color TV, and also because of their importance in such areas as pulse circuitry, instrumentation systems, and computing machines. Ideally, in the processing of signals, the changing of frequency response without changes of phase, and vice versa, would provide the most effective method of compensation. Unfortunately, however, it is almost impossible to change the frequency response without altering the phase also. Some of the methods of

controlling phase and frequency relations are discussed after an examination of the basic relations to determine the problems.

24. Linear-phase Conditions. A linear-phase circuit is one which introduces a constant delay of all frequencies which it passes. Since the delay is seldom zero, the phase shift of a constant-delay circuit should be a linear function of frequency over the entire operating range. The construction of a device capable of providing a constant delay yet, at the same time, altering the frequency characteristics of the wave being transmitted appears to be a problem similar to the one of reducing the entropy of a closed system: present knowledge seems to indicate that it may never be solved for the general case. The availability of a device capable of providing constant delay and frequency compensation simultaneously would improve the signal-to-noise ratio in pulse systems, since it would simplify the separation of a pulse in terms of its lower-frequency components.

FIG. 18. Constant-amplitude variable-phase circuit.

25. Phase-shift Networks. Two types of phase-shift devices are available to the engineer—the frequency-dependent variable-delay network, and the frequency-independent variable-delay network. Both these types of circuits find frequent use. The group of frequency-dependent variable-delay networks includes almost all the commonly used networks found in electronics. Such circuits as *RC* and *CR*, *RL*, and *RLC* all show a characteristic which has variable phase shift and variable delay if examined over a sufficiently wide range of frequencies. Only a limited number of circuits, those which have input-output relations similar to that of the circuit shown in Fig. 18, in which the output may be represented as the radius vector of a semicircle as a function of frequency, can provide controllable delay with negligible change of frequency response.

Simple circuits having a variable delay may be grouped into four basic categories, namely, those showing a net phase change of zero degrees in two frequency ranges, these ranges being separated by an area of transition; those showing 90-deg shift; those showing 180-deg change in phase shift over a narrow range of frequencies; and those showing a phase-shift difference of approximately 90 deg over a range of frequencies, but at the same time having a uniform amplitude response as a function of frequency.

26. Zero-net-phase-shift Networks. Networks which provide a net phase change of zero over two ranges of frequency but which have a phase shift other than zero between the two ranges are typified by those shown in Fig. 19. These circuits have a gain K_1 over one range of frequencies, K_2 over a second range of frequencies, with a net phase shift of zero within the ranges. Such circuits find wide application in the compensation of feedback and servoamplifiers because of the attenuation which can be introduced without at the same time introducing troublesome phase shifts, shifts sufficient in many cases to cause oscillation. The use of these circuits often permits the loop amplification of an amplifier to be increased by as much as 20 db without development of instability. In terms of components Z_1, aZ_1, and jX_2, the gains are tabulated in Table 3.

FIG. 19. Zero-degree-phase attenuation circuits.

A network having properties similar to those of Fig. 19*a* is shown in Fig. 19*b*. With the definitions indicated, the equations applying to Fig. 19*a* and *b* are identical. The equation for amplification is

$$K = [a(1 + a) + u^2 + ju]/[(1 + a)^2 + u^2] \qquad (23)$$

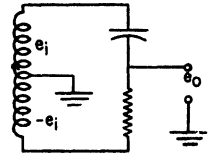



Table 8. Zero-degree Phase-shift Circuit

Frequency function	Gain	Phase, deg
$u \ll 1$	$\frac{a}{1+a}$	0
$u = \frac{X_2}{Z_1}\left(\text{or } \frac{b_2}{Y_1}\right) \approx \sqrt{1+a^2}$	$\frac{a(1+a)+u^2+ju}{(1+a^2)+u^2}$	$\frac{u}{a(1+a)+u^2}$
$u \gg 1$	1	0

Z_1 = attenuating impedance, normally a resistance
aZ_1 = loss impedance, normally a resistance
a = ratio of shunt to series resistance
X_2 = switching reactance, usually capacitive
Y_1 and aY_1 = conductances corresponding to aZ_1 and Z_1, respectively
b_2 = switching susceptance corresponding to X_2
u = parameter which is a function of frequency. It has values of X_2/Z_1 in Fig. 19b, b_2/Y_1 in Fig. 19a
ϕ = phase angle between input and output

The phase shift ϕ may be determined from

$$\tan \phi = u/[a(1 + a) + u^2] \tag{24}$$

and the maximum phase shift from

$$\phi = \arctan 1/[2\sqrt{a(1 + a)}] \tag{25}$$

27. 180-deg Circuits. Circuits capable of producing a phase shift of 180 deg over an extended range of frequencies introduce no effect on the applied wave form other than inverting it, with the result that variable delay is not introduced. Circuits which introduce approximately zero degrees phase shift over one range and 180 deg over an adjacent range with a transition between, however, do introduce a variable delay. An *RLC* circuit is capable of producing such an effect when properly excited. In such a circuit (Fig. 20) the voltage produced at the output is shifted approximately 90 deg at the frequency for which $\omega^2 LC = 1$, approximately zero degrees for frequencies appreciably less, and approximately 180 deg for frequencies appreciably greater. Such circuits are used only at very high frequencies, however, as at low frequencies simple transformers and electronic phase inverters are usually better.

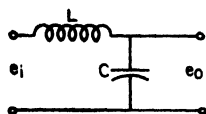



Fig. 20. *LC* filter.

28. 90-deg Circuits. Circuits, such as integrators and differentiators, which shift the phases of frequency components by approximately 90 deg when integrating or differentiating, cause a significant change in the appearance of the wave (Fig. 21). The change is due partly to the phase shift and partly to the change in amplitude response. Further data may be found in Secs. 4 and 10.

29. Wide-band 90-deg Circuits. Wide-band circuits designed to produce 90 deg difference in phase in the outputs from two channels are used extensively in some single-side-band (SSB) modulation systems. The phase-shift technique uses two audio signals which are alike except for the 90-deg phase difference, and two carrier signals which also have a 90-deg phase difference. These signals are introduced into two balanced modulator circuits (Sec. 68) (each modulator receives one of the audio signals and one of the carrier signals). After modulation the two output signals are added to eliminate either one or the other of the side bands. Reversing the polarity of one of the modulation inputs will permit shifting from one side band to the other.

A form of all-pass network normally is used in the production of wide-band phase-shift circuits as used in SSB. The basic circuit is shown in Fig. 22. A group of these circuits are made up in two chains *A* and *B* with the capacitances in the chains adjusted so that a phase angle of approximately 90 deg is developed between the

respective frequency components in the output waves from the two chains. The differential phase shifts within any pair of circuits may be between 10 and 30 deg, depending on the band width and the precision required in the phase shift. The best procedure for adjusting these circuits is to plot the phase difference as a function

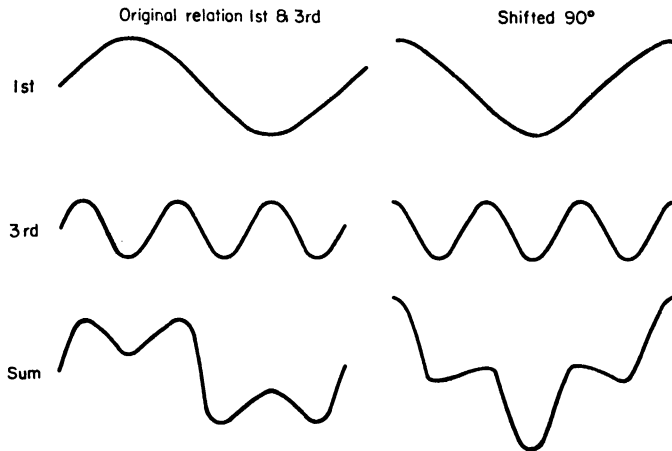



Fig. 21. Effect of phase on wave form.

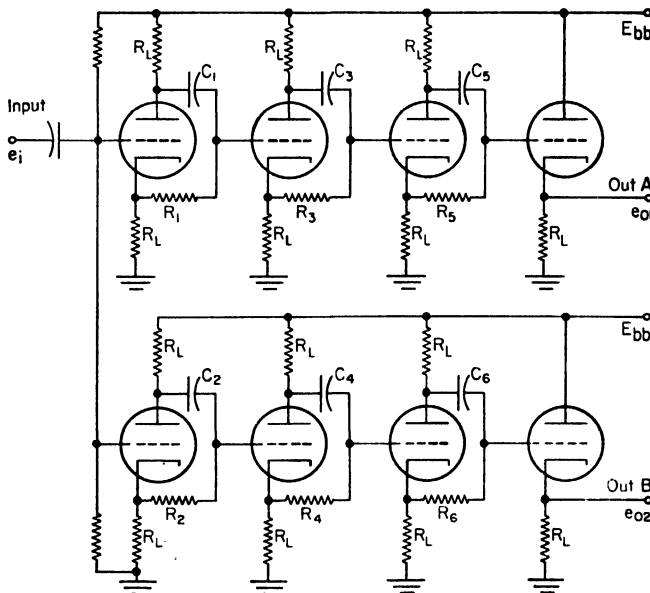



Fig. 22. Wide-band quadrature circuit.

of frequency, and then adjust the values of some of the capacitances to provide the required characteristics.

30. Minimum- and Nonminimum-phase Networks. Minimum-phase networks normally include those which are constructed from passive devices, R, L, C; however,

certain circuit configurations which do not contain active elements may possess nonminimum-phase characteristics. The principal distinguishing characteristic between the two types of circuits is in the positions of the zeros of transmission in the expression for the Laplace transform of the transfer admittance for the network; these zeros may be either on the imaginary axis or in the right-half plane for a network with nonminimum phase. Wide-band all-pass networks are possibly one of the best examples of nonminimum-phase networks.

All-pass networks have a set of transmission zeros in the right-half plane at points which are mirror images with respect to the imaginary axis of the poles of the Laplace transform of the network transfer function. The result is that the network passes all frequencies with a constant signal amplitude, and introduces phase shifts as a function of frequency which can be used for compensating for undesired phase shifts.

31. Conditions for Constant Delay. Constant delay may be obtained with most types of coupling circuits if they are used in the appropriate ranges. Since the maintenance of constant delay is frequently of great importance the next few paragraphs discuss the conditions required to obtain constant delay in some of the more common circuits. More information may be found in books and articles on wave filters and delay lines, on telephone and television circuits, and on similar subjects.

Constant delay with RC and RL circuits may be obtained only over a narrow range of frequencies for which the amplification as a function of frequency is almost exactly constant. The region of constant delay extends to a maximum phase shift of less than 10 deg, or a change of amplification of approximately 2 per cent. For low-pass circuits, the maximum frequency to which an approximately constant delay may be obtained is approximately one-sixth of the frequency for which $R = X$, and for high-pass circuits, the minimum frequency is approximately six times that at which $R = X$. A factor of either 10 or one-tenth is normally used in the design of television circuits to make certain that the phase and flatness specifications are fulfilled.

R-f and i-f circuits are also subject to variable delay. Depending on the number of amplifier and selection stages being cascaded when constant delay is required, the total phase shift permitted per stage may range from 1 to 10 deg per selection circuit, depending on the phase accuracy required and on the total number of amplifiers and selective circuits in the over-all system. The error per selective circuit may be approximated by adding the successive differences between the phase angle in radians and the sine of that phase angle for each successive circuit involved. In other words, the error for one stage is

$$E \approx \theta^3/6 \tag{26}$$

where θ is the phase angle in radians. For n stages, the equation may be written

$$E \approx \sum_{i=1}^{n} \theta_i^3/6 \tag{27}$$

The differential delay for television signals should be kept to a fraction of a microsecond if a clear and accurate picture is desired. For color, true color registry requires that the differential delay be less than 50 deg at 3.85 Mc, or less than 0.1 μsec.

32. Variable Delay in Transmission Lines. Any type of transmission line or delay line is subject to delay distortion if the effective propagation velocity varies with frequency. The conditions for constant delay in any transmission line are that, on a per-unit-length basis, the following relation between the parameters exists:

$$R/L = G/C \tag{28}$$

where R and L are the series R and L per unit length, and G and C are the shunt conductance and C per unit length. Further data on this problem may be found in any of the many texts on design of telephone circuits and on design of wave filters.

33. Practical Effects of Variable Delay. The maintenance of constant delay is of much greater importance than the maintenance of a uniform frequency-response char-

acteristic in many electronic systems. The variation of delay with frequency can excite ringing within a circuit and can produce noticeable transients and changes of appearance and timbre. Circuits which are compensated to give a flat frequency response tend to overshoot and to ring, whereas circuits which are compensated to a flat phase (constant delay) characteristic do not. A phase-compensated r-f or i-f amplifier system has a monotonic decay of amplification on either side of the peak frequency (a Bessel[1] response), rather than an equal ripple or related type of response across the pass band.

Constant delay is important in audio as well as video applications. DeRosa[2] has shown by synthesizing tones that the timbre of a combination is dependent on the phase relations of the harmonics. In addition, the phase transients can produce chirps and buzzing or rattling often noticed in poorly designed high-fidelity equipment.

Systems which use the comparison of the phase of sine-wave signals are highly sensitive to the effects of variable delay, particularly if a combination of several frequencies is used in the measurement. The phase error resulting from each individual measurement must be kept small enough so that there is no ambiguity in reading one phase with respect to another. (In other words, the error on one frequency should be less than one-third of the phase shift corresponding to one cycle of phase of the next higher frequency.) Often it is possible to get the required accuracy in the circuitry, but errors due to propagation vagaries or multipath effects may be large enough to render a phase-measuring system of little use.

Variable delay in the r-f and i-f systems for conventional a-m communication systems seldom introduces difficulty because of the redundancy of the information being transmitted. The variation of the delay has the effect of introducing frequency modulation into the signal and reducing the over-all percentage of modulation. In addition to the reduction of the percentage of modulation, the variation of delay introduces distortion into the modulation signal.

A similar conclusion may readily be drawn for the effect of variable delay on an f-m communication system. The band width normally used with f-m systems is considerably wider than that used with a.m., with the result that internal-delay problems may be somewhat more serious than with a.m. Variable delay once again reduces the modulation index and introduces distortion as a result of conversion of part of the f-m signal to a.m.

The effect of variable delay on pulse-modulated signals has largely been neglected in the literature. If the Fourier spectrum of a pulse is examined, however, it can be shown to be the summation of a small amount of each of a long series of harmonically related components. Since the form of the pulse is affected by comparatively small changes in the phase relations of these harmonic components, variable delay can be very important, particularly as the duty-cycle μ is reduced. The equation for the wave form of a pulse in terms of its harmonics may be written

$$E_{(x)} = k\left(\mu + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{\sin \pi n\mu \cos nx}{n}\right) \tag{29}$$

where x is the radian angle corresponding to time delay within the period of repetition, n is the order of the harmonic, and the pulse is symmetric about the point $x = 0$. Since the ratio $[(\sin \pi n\mu)/\pi n\mu]$ has the approximate value of unity for all values of n such that $n\mu$ is less than $\pi/6$, for values of μ of approximately 0.01 the magnitudes of each of the harmonic components up to the fiftieth are nearly equal. Evidently, small errors in the phase of the different components of the wave can cause considerable loss of recognizability in pulsed signals, particularly when the duty cycle is short.

Variations in the delay in circuits, particularly in systems which utilize large amounts of feedback, can easily make possible oscillation at a frequency which is

[1] Weinberg, L., Modern Synthesis Network Design from Tables, *Electronic Design*, **4** (18–21), Sept. 15, Oct. 1, Oct. 15, and Nov. 1, 1956.
[2] DeRosa, L. A., Phase Distortion in Audio Systems, IRE National Meeting, 1948.

determined by the delay. Conventional open-loop tests on a circuit will show the existence of critical delay conditions only if the elements which provide the abnormal delay are part of the loop in the form in which it is tested. The introduction of a loss network of the zero-degree-phase type can reduce the loop gain at the critical frequency, thereby stabilizing the system.

CIRCUITS WITH AMPLITUDE NONLINEARITY

34. Basic Operation. The discussion in the following sections, with the exception of Secs. 79 to 85, is directed to applications in which the variation of the derivative of the output current with respect to the input voltage (the transfer function) is intentionally made large to take advantage of the resulting multiplicative properties of the circuit. If, for example, the function relating the output current to the input voltage,

$$I = F(E) \tag{30}$$

is expanded in a Taylor series about the point E_0, and the variable $E - E_0$ is taken to be e_i and the variable $I - I_0$ to be i_0, it may be written as

$$i_0 = F'(E_0)e_i + F''(E_0)e_i^2 + \cdots \tag{31}$$

The first term in this power-series expansion expresses the conditions for linear operation at the point (E_0, I_0), whereas the remaining or nonlinear terms express what might be called the multiplicative properties of the device.

35. Multiplicative Action and LTV Circuits. Multiplicative action may be obtained in at least one other way, namely, by making the function F' variable under the control of some voltage other than E_0. If F' is such a function, then its value is dependent on the variation of an external parameter E_c, giving, when e_i is sufficiently small, a circuit whose behavior is linear for any fixed value of E_c, but whose properties change markedly with the magnitude of E_c. Such a circuit may be called a *linear-time-variant* (LTV) circuit.

In many applications the two types of circuits, namely, nonlinear and linear-time-variant, may be used interchangeably. Each, however, has its own special advantages. Since most time-variant circuits can be made to behave as nonlinear circuits by substituting e_i for E_c, the circuitry for all is discussed under nonlinear circuits, and the special advantages of time-variant circuits are described separately.

USES OF NONLINEAR CIRCUITS

Nonlinear circuits are especially useful in the following applications which are commonly encountered in electronics:

1. Detectors
2. Limiters and clamping circuits
3. Mixers
4. F-m detectors
5. Modulators
6. Pulse shaping and forming circuits
7. AVC circuits

In each of these applications, nonlinearity is intentionally introduced to make possible the required action. Since distortion and nonlinearity are essential to these devices, some data on the calculation of the characteristics of nonlinear devices are included in Secs. 79 to 85.

36. Detection. Detection is the process of removing desired information from a composite signal which conveys it through a communication or a processing system. The system may be a radio receiver of any type, a chopper-stabilized amplifier or power supply, a servo system, or any other of a variety of devices. The conventional a-m detector is discussed first to show how nonlinearity brings about the desired changes.

The amplitude-modulation detector separates the intelligence conveyed by the

variation of the amplitude of a carrier from the carrier itself through the use of the nonlinearity of the amplitude response of the detector. As was shown in Sec. 34, the output of a network may be expressed in terms of its input through the equation

$$i_0 = F'(E_0)e_i + F''(E_0)e_i^2 + F'''(E_0)e_i^3 + \cdots \tag{31}$$

The derivatives of the function F may be replaced by the conductance and its derivatives:

$$F^{(n)} = g_{(n-1)} \tag{32}$$

If, then, a voltage of the form $E\left(1 + \sum_{i=1}^{n} K_i \sin \omega_{mi}t\right) \sin \omega_c t$ is substituted for e_i, the output current, neglecting carrier frequency components (ω_c and $2\omega_c$), is

$$i_0 = E^2 g_{m1} \left\{1 - \sum_i^n K_i \sin \omega_{mi}t + \sum_i^n \sum_j^n K_i K_j [\cos (\omega_{mi} - \omega_{mj})t - \cos (\omega_{mi} + \omega_{mj})t]\right\} \tag{33}$$

Since the double-summation term represents the distortion and intermodulation terms, the square-law detector does not have satisfactory characteristics for received signals having values of K_i greater than 0.1. An ideal detector will provide rectification without introducing a double-summation term.

37. The Piecewise-linear Detector. One form of detector which is free of the double-summation term is known as the piecewise-linear detector, a detector which, with positive signal voltage, has one fixed value of g and, for negative input, a different value, with a sharp change of the value of g occurring between the two states when the signal voltage approaches zero (Fig. 23). With such a detector, more current is passed per volt of applied signal with one polarity than the other, with the result that the average d-c current after passage through the device is no longer zero.

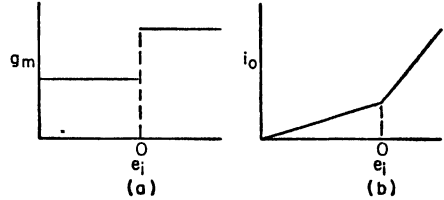



FIG. 23. Piecewise-linear circuit.

Since the higher-order terms g_j, where j is greater than unity, can and should be zero, no multiple-summation terms are present, and the intermodulation distortion is reduced. The diode detector is the simplest example of a device having piecewise-linear characteristics for large input signals.

38. The Diode Detector. The diode detector makes use of the unidirectional, or unilateral, characteristics of crystal and vacuum diodes to produce an output voltage which is proportional to the modulation level. The operating equation may be written as

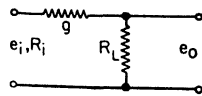



FIG. 24. Equivalent diode circuit, series diode.

$$i_0 = [g/(1 + gR_L)]e_i \tag{34}$$

where g is the conductivity of the device and R_L the terminating impedance (Fig. 24). Assuming that the two values of conductance are g_f and g_r and solving for the value of R_L which will give the maximum output gives

$$R_L = 1/\sqrt{g_f g_r} \tag{35}$$

If the signal source has appreciable internal impedance, it should be included in g_f and g_r, giving

$$R_L = 1/\sqrt{(1 + g_f R_i)(1 + g_r R_i)/g_f g_r} \tag{36}$$

where R_i is the internal impedance of the source. When the two products in the numerator of the function under the radical are large compared with unity, the required value of R_L exactly equals R_i, indicating that the required load impedance equals the input impedance, and the diode may be ignored from the circuit point of view. Consequently, at least g_rR_i, and if possible both g_fR_i and g_rR_i, should be small compared with unity.

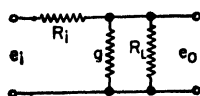



Fig. 25. Equivalent diode circuit, shunt diode.

When the source impedance R_i is sufficiently large (constant-current conditions) no rectification is obtained with the series diode; the diode then should be connected in shunt, and the source impedance should be large compared with the load impedance (Fig. 25). The correct value of load impedance is given by Eq. (35).

39. The Push-Pull Diode Detector. The push-pull diode detector (Fig. 26) generates appreciably less harmonic distortion than the single diode because of the full-wave rectification obtained. It is particularly valuable when low distortion is required in extracting the modulation from a low-frequency i-f signal. The full-wave rectification in effect doubles the carrier frequency, making the distortion half or less than that otherwise introduced, partly because of the higher critical frequency which may be selected for the output low-pass filter, and partly because of the larger number of data points defining the output signal. The design of a push-pull detector is based on the same equations as the single-diode detector, the only difference in the circuitry being the center-tapped transformer required to provide the input signal.

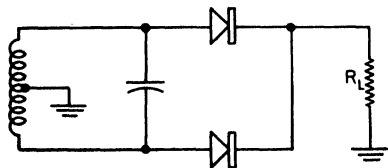



Fig. 26. Full-wave diode circuit.

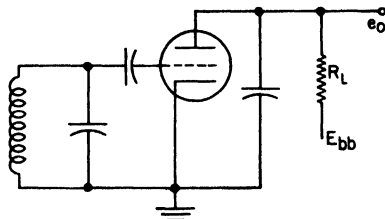



Fig. 27. Grid-leak detector.

40. The Grid-leak Detector. The grid-leak detector (Fig. 27) has, under small-signal conditions, a characteristic which is essentially piecewise linear, one in which one section is nearly of zero length, with the other having an amplitude approximately equal to the peak-to-peak amplitude of the signal being detected. The portion of approximately zero length introduces clamping, making possible very efficient detection of weak signals through the amplification of the main portion of the signal. Since, however, the linearity of a tube used as a grid-leak detector is poor if the input signal has an amplitude over 0.2 to 0.3 volt, this detector is seldom used in present-day equipment.

The grid-leak detector clamps the applied signal by grid rectification of the most positive portion of the signal, thereby producing an average bias which is approximately equal to half the peak-to-peak signal amplitude. This average bias varies with the modulation or the over-all instantaneous amplitude of the modulated wave, producing a changing average plate current in accordance with the original modulation. Amplification and distortion may be estimated either by standard techniques or by conductance techniques (Sec. 85).

41. The Power Detector. The power or plate detector (Fig. 28*a*) uses a tube operated near plate-current cutoff to take advantage of the approximate square-law characteristics of the tube. It has fallen into disuse except for applications in which an approximate power or rms voltage measurement is required.

The suitability of any tube for use as a square-law detector may be determined by examining the linearity of the variation of the small-signal amplification from cutoff

to zero bias along the load line used. The plot of amplification as a function of bias is better than a plot of output current because of the tendency of the plot of current vs. bias to integrate or average out the variations (Fig. 28*b*).

42. The Infinite-impedance Detector. The infinite-impedance detector provides linearity at least equal to that of the diode detector. Because of the large amount of inherent degeneration, an almost ideal piecewise-linear characteristic is available in the infinite-impedance detector (Fig. 29). As can be seen from Fig. 30, the correct bias for proper operation of this circuit is that value which establishes the static operating point at the mid-point of the transition region; otherwise serious distortion may be introduced with high percentages of modulation. Since the correct bias is almost equal to the value required to cut off the plate current, a fixed bias of

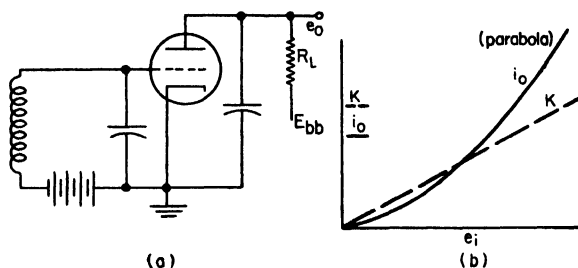



FIG. 28. Power (square-law) detector.

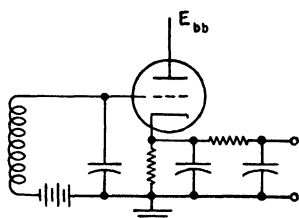



FIG. 29. Infinite-impedance detector.

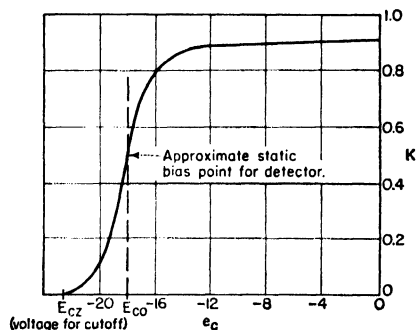



FIG. 30. Gain of infinite-impedance detector.

the required magnitude should be introduced experimentally for each tube if accuracy of rectification is of critical importance. The approximate value may be found from the plate-characteristic curves for the tube used.

The output capacitance used with the infinite-impedance detector is used to smooth the output modulation and to cut off the cathode current of the tube during all but the peak of the conduction cycle. (This cutoff phenomenon is familiar to users of cathode followers; it appears in pulse and video circuitry as the storage effect which blocks current flow during negative-going pulses.) The charge in the capacitance must decay, between pulses of current, to a value which is less than the peak value of voltage developed on the next cycle of the carrier if trace distortion is avoided; the distortion causes a loss of high-frequency response. An *RC* integrator, shown in Fig. 29, is usually used after the detector to smooth out any remaining high-frequency variations. The critical frequency for the integrator circuit should be about 100,000 cps.

The main practical disadvantage of the infinite-impedance detector is the fact that

it cannot be used to develop a control signal voltage for AVC or AGC use. For this reason, a separate AGC rectifier is usually used. It is normally isolated by a separate amplifier to prevent the variable loading of the AGC detector from distorting the signal supplied to the signal detector.

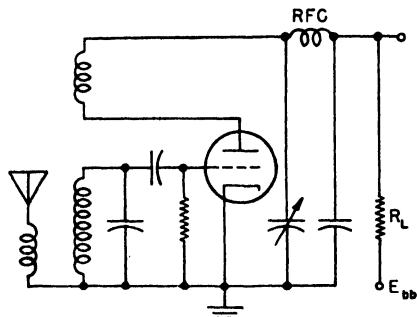



FIG. 31. Regenerative detector.

43. The Regenerative Detector. The regenerative detector, a special form of the grid-leak detector, differs from its companion primarily in the presence of a feedback circuit which makes the detector an oscillator producing weak oscillations as well (Fig. 31). This feedback circuit, in developing weak oscillations, introduces a reference carrier for the incoming signal to be differenced or mixed with, and also supplies the losses of the tuned circuit, permitting it to appear to have an extremely high Q and, as a result, to build up the incoming signal by a factor which may be as high as a million times.

The oscillation-stabilizing circuit of this detector is adjusted to have a critical frequency which is in the low audio range so that the beat between the incoming carrier and the local oscillation may be observed. (The loading is obtained from

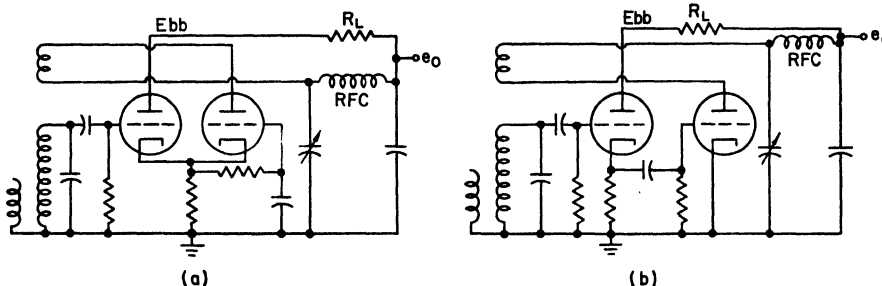



FIG. 32. Special regenerative detectors.

diode loading in the grid circuit.) Often this critical frequency may be as low as or lower than 500 cps, thereby making possible values of effective Q of the tuned circuit as high as 30,000 or more. The adjustment of the feedback to give oscillation without pulsing or *squegging* is very critical since the time constant of the grid-leak circuit is very long compared with the decay time of the tuned circuit ($2\pi Q/\omega$).

Two improved forms of regenerative detectors are shown in Fig. 32*a* and *b*. The operation of these detectors is identical with that of the more conventional type except in one respect—the diode loading is not coupled directly into the tuned circuit but is isolated from it by a low-gain amplifier or cathode follower. The setting of the regenerative coupling is somewhat more critical because of the absence of the variable diode loading on the tuned circuit; the control is accomplished by the variation of transconductance instead. The absence of the variable loading on the tuned circuit greatly enhances the stability of the detector and makes its use on a-m signals practical. Otherwise, the feedback or regeneration must be reduced to a setting which is just below oscillation with a-m signals.

44. Limiting and Clamping. Devices having nonlinear amplitude characteristics as a function of applied voltage may make excellent limiters and clamping circuits. Ideally, a clamping circuit normally cuts or clamps a signal at a certain voltage level and for one polarity, whereas a limiter clips on both polarities at predetermined voltage

levels. As with detectors, devices which possess piecewise-linear characteristics make the best limiting and clamping elements, particularly if one or two of the sections have zero gain for input signals within their control ranges, with the remaining linear area providing a constant gain.

The characteristics of a signal after passage through a clamping circuit are dependent on the amplitude of the applied signal unless the clamping circuit has ideal piecewise-linear characteristics with the break point between the linear sections occurring at the zero level of the input signal. Since ideal conditions are hard to achieve in practice, the applied voltage should be of sufficiently high amplitude that any anticipated variations will not introduce difficulties. (Consequently d-c restoration or clamping in TV circuits is usually accomplished at the picture-tube grid or cathode.) Figure 33 shows the result of clamping a sine wave.

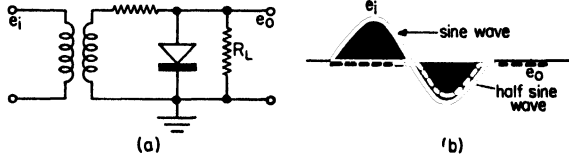



FIG. 33. Circuit passing a half wave.

45. Limiting. Limiting is usually used to minimize the effect of noise energy on the operation of noise-sensitive circuits. In a sine-wave signal, the voltage actually remains in the neighborhood of the peak value much longer than it remains near zero, with the result that small variations due to noise may be suppressed by limiting the amplitude. The clamping minimizes the variation in the energy per half cycle in the signal, taking out both noise and amplitude modulation. As can be seen in Fig. 34, the only noise which can affect the received signal in a heavily limited signal is that which occurs during the transition periods.

FIG. 34. Effect of limiter on signal with noise.

The total energy per half cycle of a voltage wave may be found by integrating the square of the output voltage over the period of the wave. When the signal voltage is less than the limiting level, the value follows the integral of a sine-squared function, whereas, when the voltage is greater than the limiting level, the value is the integral of the limiting value. The total energy per cycle then is proportional to

$$P = (a^2 - b^2) \arcsin (b/a) + \pi b^2 - b \sqrt{a^2 - b^2} \qquad (37)$$

where the amplitude of the sine wave a is equal to or greater than the limiting level b. In case a is infinite, the first and third terms cancel, leaving $P = \pi b^2$; in case $a = b$, $P = \pi b^2/2$. Consequently, such a limiting circuit restricts the total energy in a limited wave to a 2:1 ratio, or a voltage ratio of $\sqrt{2}$:1. Additional limiters can be used to restrict the voltage variation even further.

For applications in which phase stability is important, limiting should be accomplished by the use of a circuit which is designed to clamp symmetrically on both polarities. Otherwise, the tuned circuits which are excited by the voltage and current pulses may introduce a phase shift as a result of the dissymmetry. For ordinary communications use a transient of little importance would be introduced; in precision instrumentation, however, the transient could be troublesome.

46. The Diode Clamp Circuit. The single diode, as in Fig. 35*a*, may be used as a clamping device or a rectifier. Its characteristics depend primarily on its mode of operation; if it is used to couple a source and a load both of which have either reactive

characteristics (of the same polarity) and negligible d-c resistance or resistive characteristics and negligible reactance, it will restrict the output at a selected value of voltage. Otherwise, the reactive element or elements can store energy and discharge it slowly, causing clamping to occur at a voltage level which is a function of input voltage.

The diode, as normally used, distorts the applied signal voltage, thereby generating harmonics of the applied signal. For a sine-wave voltage clamped at the zero voltage line, the harmonics generated have the relative amplitudes

$$A_n = 2/(n^2 - 1) \qquad (n \text{ even}) \tag{38}$$

Since the amplitudes of the harmonics are nearly inversely proportional to the square of the harmonic order, the harmonic generating efficiency of a clamp circuit designed

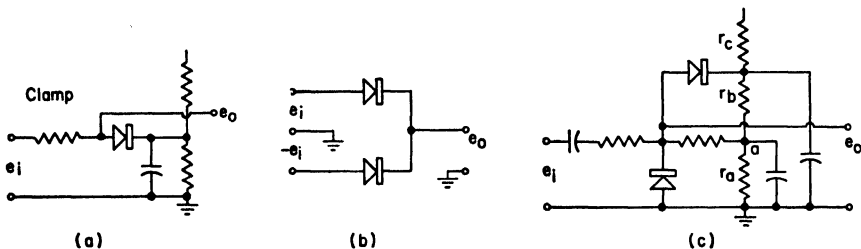



Fig. 35. Clamping and limiting circuits.

to clamp at the zero voltage point is very poor. Efficiency may be improved by shortening the conduction cycle, giving the relative amplitudes values in accordance with

$$A_n = K \left\{ \frac{C \sin [(n-1) \arccos (a/c)]}{n-1} + \frac{C \sin [(n+1) \arccos (a/c)]}{n+1} - \frac{2a \sin [n \arccos (a/c)]}{n} \right\} \tag{39}$$

where a is the bias voltage applied and c is the amplitude of the applied sine wave [the conduction angle is plus or minus arccos (a/c)]. When a is nearly equal to c, A_n is very similar to the corresponding coefficient for the Fourier spectrum of a rectangular pulse of approximately equal duration.

A pair of diodes may be connected in a balanced circuit as shown in Fig. 35*b* to provide either full-wave rectification or frequency doubling. A low-pass output circuit is used with a full-wave rectifier, and a circuit is tuned to the second harmonic when the circuit is used as a frequency doubler. The input voltage should be balanced with respect to ground with the circuit in Fig. 35*b*; otherwise either unsatisfactory smoothing or unsatisfactory doubling will result.

A circuit having two diodes connected as in Fig. 35*c* may be used as a limiter; the two diodes are both biased to be nonconducting for small values of signal voltage. The bias circuit shown in the figure (consisting of r_a, r_b, and r_c) should be adjusted to place the conduction voltage level of one diode as much positive with respect to the reference point a as the other conduction level is negative. Then the limiting achieved is symmetrical, and maximum operational stability for the limiter should result. The current flowing in the bleeder circuit should be large compared with the peak diode current expected (at least twice the peak value) to make certain that the limiting levels do not change with applied signal voltage. The dual-diode limiter has one major disadvantage—it has an amplification which is appreciably less than unity. Where higher amplification is required, the gated-beam limiter and the cathode-coupled limiter should be considered.

47. Triode and Pentode Clamp Circuits. Triode and pentode tubes are often used in clamping amplifiers by taking advantage of the clipping action which can be obtained

from the grid conduction in the positive-bias area. As shown in Fig. 36, a series resistance R_d may be introduced in series with the grid, or the internal resistance in the source itself may be used to limit the grid current and provide the clamping; the value of resistance selected should be large enough that the product of the input conductance of the grid (g_c or g_{11}) and R_d is large compared with unity and g_c or g_{11} is the input or base conductance for tubes or transistors, respectively.

FIG. 36. Triode clamp circuit.

The existence of clamping due to cutoff in triode and pentode tubes makes them useful in limiters as well as clamping circuits. Of the two tube types, the pentode usually is somewhat more effective because of the sharper cutoff obtainable. The range of screen voltages which may be used with pentodes makes available a gap between grid conduction and current cutoff which may be adjusted to suit the convenience of the user. The load impedance may be selected to take advantage of the characteristics of the individual tube. Triode and pentode clamping circuits may be designed to function at carrier frequency by the use of a tuned load impedance or at modulation frequency by the use of a load resistance alone.

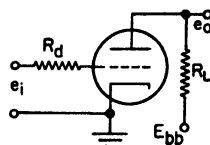
In the use of either a triode or a pentode as a limiter, the bias selected for the control grid should be that which corresponds to a plate current half that for zero bias. Since the symmetry of operation is rather poor otherwise, the introduction of some cathode degeneration may improve the operating characteristics if the operating frequency is sufficiently low (usually less than 1 Mc). Degeneration increases the amount of grid-voltage change required to achieve limiting, with the result that an adjustment of either the input signal amplitude or, with a pentode tube, the screen voltage, or both, may be required to make the limiting effective. The peak-to-peak voltage at which limiting commences may be found from the equation[1]

$$E_t = I_{kp}R_k - E_{cn} \tag{40}$$

where E_t is the peak-to-peak voltage at which limiting commences, I_{kp} is the total cathode current at zero bias, E_{cn} is the approximate bias required to cut off the plate current in the tube, and R_k is the resistance between cathode and ground.

[1] Standard IRE definitions are small letters as symbols along with a single subscript for small-signal or instantaneous currents, voltages, and conductances or resistances. For fixed values at specific operating points, such as at the static operating point with or without signal, however, a capital-letter symbol is used along with two or more identifying subscripts. The following table indicates some additional definitions used in this chapter:

Instantaneous symbol	Static no-signal	Static with-signal	Grid-bias limit: Positive	Grid-bias limit: Negative
i_b	I_b	I_{bs}	I_{bp}	I_{bn}
e_c	E_c	E_{cs}	E_{cp}	E_{cn}
e_b	E_b	E_{bs}	E_{bp}	E_{bn}
g_m	G_m	G_{ms}	G_{mp}	G_{mn}
g_p	G_p	G_{ps}	G_{pp}	G_{pn}
g_{mi}	G_{mi}	G_{mis}	G_{mip}	G_{min}
K	K	K_s	K_p	K_n

The only symbol above which is somewhat irregular, K, is so because of the standard use of K for small-signal gain. Consequently, the required subscripts are added directly. As can be seen, the additional symbols for the parameters at the maximum positive grid bias and the most negative grid bias are made by substituting either a p or an n, respectively, for the s used to specify the static with-signal conditions. In this way, the characteristics at the bias limits may be specified, simplifying distortion and mixer-design calculations.

48. The Cathode-coupled Limiter. The cathode-coupled limiter (Fig. 37) makes use of a cathode follower and a grounded-grid amplifier connected in cascade; the first of these tubes acts as a repeater and at one limit shuts off the plate current in the output amplifier, and at the other limit has its own current shut off by the action of the amplifier. Because limiting is accomplished by the cutoff of the plate current in either the one or the other tube, rather good symmetry can be obtained if the bias is adjusted to place the static operating point of the amplifier at the mid-point of the output characteristic (Fig. 38). The cathode-coupled limiter has the advantage of having a small-signal amplification in the active range between 5 and 20. It is normally used at frequencies below 10 Mc, and the balanced diode limiter above 10 Mc.

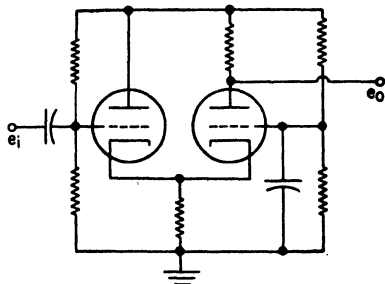



FIG. 37. Cathode-coupled limiter.

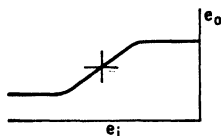



FIG. 38. Input-output relation, cathode-coupled limiter.

49. The Gated-beam Limiter. The gated-beam limiter is unique in that the tube used has a transconductance which varies approximately symmetrically about the mid-point of its useful bias range, namely, between zero bias and collector cutoff bias. This phenomenon is shown by the curves in Fig. 39 for the 6BN6 tube. When this tube is used as a limiter, the quadrature grid is grounded, and the input bias is placed at the value required to produce maximum static transconductance. The limiting action occurs in the tube both with the quadrature grid grounded and also with either external or self- (negative capacitance coupled) excitation. The screen or accelerator voltage for this tube may be selected to give the width of conduction gap desired; the value chosen should be small enough that the power dissipation of the various electrodes is well within their ratings and at the same time large enough to give an adequate collector output signal.

50. Mixers. Electronic mixers are multipliers; they generate sum and difference frequencies by the multiplication, transforming the modulation carried on a received signal onto the frequencies generated by the multiplicative action. Mixing or multiplication of voltages is accomplished by causing one voltage to vary the amplification experienced by the other in passing through the multiplier. The equation controlling this action is

$$K = f(e_r) + e_s f'(e_r) \qquad \text{or} \qquad e_0 = \int f(e_r) + e_s f(e_r) \tag{41}$$

where e_r is the reference oscillator voltage, e_s is the signal voltage, with e_r much larger than e_s, and $f(e_r)$ expresses the variation of the voltage amplification of the multiplier as a function of the reference voltage. The amplification function $f(e_r)$ may be written in power-series form as

$$f(e_r) = K_0 + K_1 x + K_2 x^2 + K_3 x^3 + \cdots \tag{42}$$

where x represents the ratio of the instantaneous value of e_r to the peak value of e_r, e_{r1}, and the Ks are the coefficients of the expansion.

The conversion amplifications for some of the more commonly used conditions of operation are listed in Table 4.

When the input and the output frequencies are low enough that values of K_p greater than 2 or 3 may be obtained with a triode or a multigrid tube, then the mixing

resulting from a linear variation of amplification is usually chosen in preference to the mixing obtained from a diode under piecewise-linear operating conditions. Since diode and triode mixers have lower inherent noise levels than multigrid mixers, they require much less signal amplification ahead of the mixer; unless, however, the locally generated oscillator signal is adequately isolated to minimize detuning with diode and

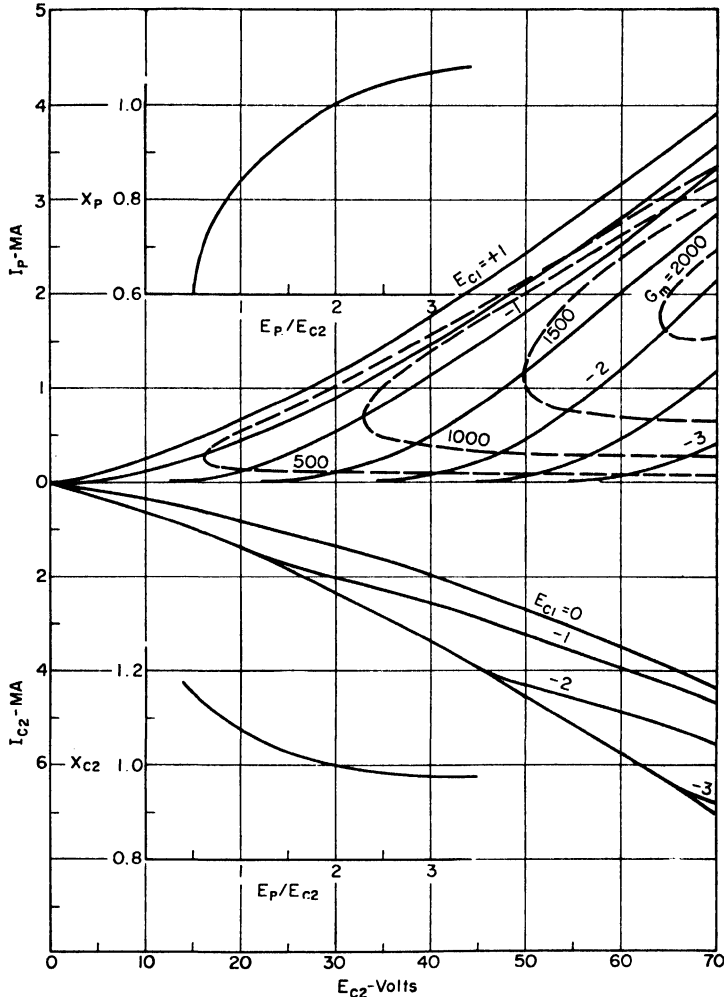



FIG. 39. Characteristics, 6BN6.

triode mixers, poor stability may result. The multigrid mixer has been used extensively for low-frequency receivers because of the greatly reduced coupling between the input and local oscillator circuits.

In applications requiring the precise determination of carrier phase, even multigrid mixers may have sufficient coupling to give unstable operation. Stabilization may be obtained for such applications by the use of low-impedance input circuits (a broad-

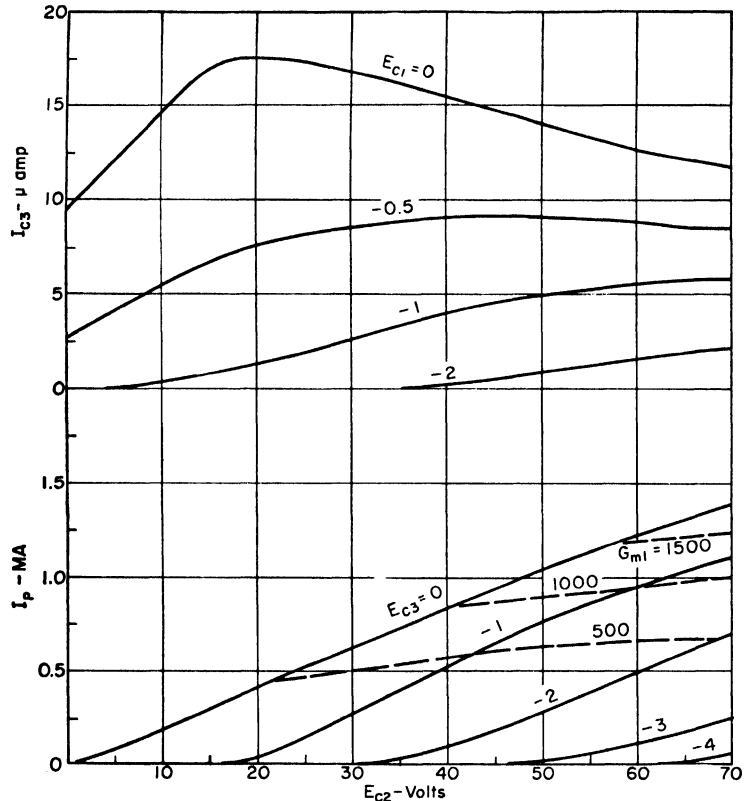



FIG. 39a. Quadrature grid-current and converter curves, 6BN6, R_K = 300 ohms.

Table 4. Conversion Amplification

Type of circuit*	Type of output	Conversion gain
Piecewise linear $K = K_p$ when e_i 0 $K = K_n$ when e_i 0	Average Peak	$\frac{K_p - K_n}{\pi}$ $\frac{K_p - K_n}{2}$
Piecewise linear $K = 0$ when $e_i \geq e_r$ $K = K_n$ when $e_i < e_r$ and $-e_r \leq e_i \leq e_r$	Average Peak	$\frac{2K_n}{\pi}$ K_n
Quadratic $-K = K_0 + K_1 \frac{e_i}{e_r}$ where $-e_r \leq e_i \leq e_r$	Peak	$\frac{K_1}{2}$
In above, taking $K_p = K_0 + K_1$ $K_n = K_0 - K_1$	Peak	$\frac{K_p - K_n}{4}$

* See Note in Sec. 47.

band 50-ohm termination on both the input grids is ideal) and the use of by-pass capacitors on the screen and cathode which provide adequate by-pass action to a frequency which is at most a tenth of the mixer output frequency.

51. The Cathode-coupled Mixer. The cathode-coupled circuit shown in Fig. 40 can be used to provide mixing action with low noise and negligible interaction between the inputs. This circuit is of growing importance because of its inherent stability coupled with its low noise figure. The magnitude of its conversion amplification, since it is dependent on the linear term in the function $f(e_r)$, may be evaluated in terms of the coefficient K_1 of Eq. (42). As long as f can be represented in terms of a quadratic expression in (e_r), then $K_1 = (K_p - K_n)/2$ and the conversion amplification is

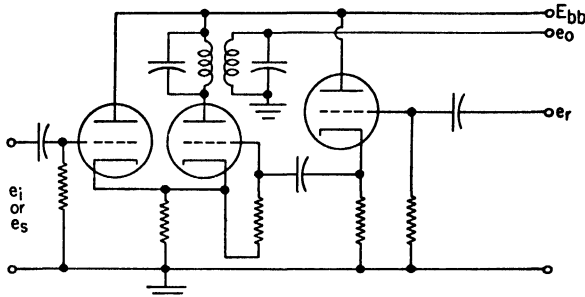



FIG. 40. Cathode-coupled mixer.

$$K_c = (K_p - K_n)/4 = K_1/2 \quad (43)$$

52. The Triode Mixer-design Technique. The simple triode mixer may be used at very high frequencies in applications in which the difference, or i-f, frequency is high. Under such conditions, the efficiency of the cathode-coupled mixer may be low. The triode mixer is commonly used in television and u-h-f communications receivers.

The calculation of the conversion amplification of a practical triode mixer requires the use of the small-signal data on the characteristics of the active device over the full range of operation. Because the mixer is designed to suppress the incoming signal frequencies and pass only the difference frequency, it has an output impedance approaching zero at the two input frequencies, and a finite output impedance (between 1,000 and 100,000 ohms) only at the difference frequency. For this reason, the input load lines R_{Li} in Fig. 41 are nearly vertical, and the family of output load lines R_{Lo} has a slope corresponding to the i-f output impedance. A set of output load lines is required because the local oscillator voltage in effect changes the operating point of the mixer amplifier instant by instant.

The basic amplification data may be calculated at a series of points which are at the intersections of the input load lines with respective typical output load lines for different values of local oscillator voltage. The output load lines selected from the family usually are chosen to represent equal increments of reference signal voltage between the negative and positive limit values. A value of transconductance and a value of plate conductance are read at each intersection, and the data at each intersection are inserted in the equation

$$K = -g_m R_L/(1 + g_p R_L) \quad (44)$$

where R_L is the tuned impedance at the output frequency. The conversion gain may then be determined from the appropriate equation in Table 4.

The local oscillator voltage should be large enough to shift the grid bias of the tube from approximately zero to the cutoff bias along the input load line. The exact value of the voltage is not critical as long as it is large enough to reduce the amplification obtained on the negative bias limit (E_{cn}) to a value which is less than a quarter of the value at the positive limit bias voltage (E_{cp}). The use of grid-leak rectification is recommended to provide the proper static bias and the correct operating conditions for the control grid of the triode mixer.

53. Multigrid-mixer Design. The design of the multigrid mixer circuit differs from that for the triode mixer primarily because of the use of separate grids for the introduction of the incoming signal voltage and the local oscillator voltage. The data required for circuit design, as with triodes, include static data on the variation of plate current with screen voltage at the various values of grid bias, and also data on transconductance as a function of the same parameters. These data may be organized to provide, first, the information required for the design of the mixer as an amplifier and, second, to permit a modification of the design to provide the required mixer characteristics. The first step, the design of an amplifier, may be accomplished by the use of the manufacturer's recommended operating conditions or by the use of conductance

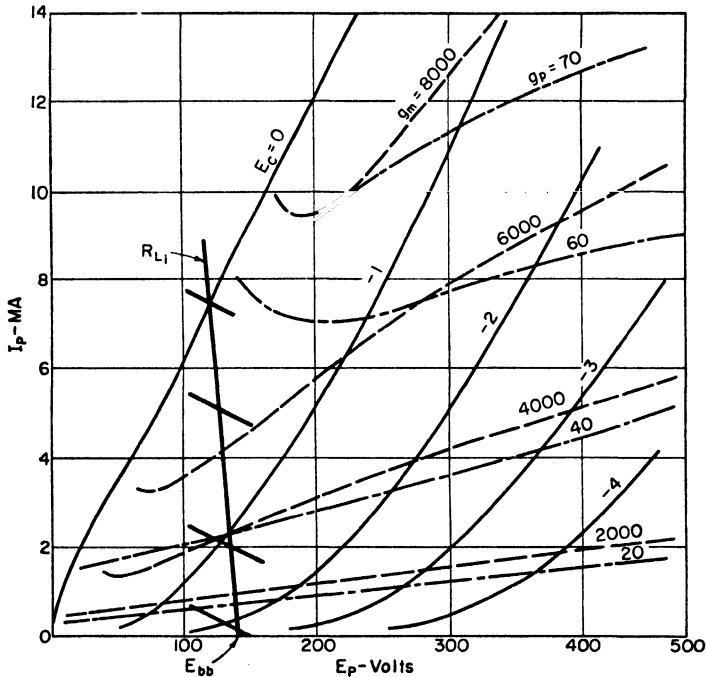



FIG. 41. Plate characteristics, 12BZ7.

techniques.[1] The second step, calculation of the converter characteristics, is readily accomplished by a special set of conductance curves called screen-converter curves. These curves (Fig. 42*b* and *c*) include two sets, one showing the variation of plate current with screen voltage for constant values of bias on the first grid along with contours of constant transconductance for the third grid, and the second set showing the variation of plate current for constant values of bias on the third grid along with contours of constant transconductance for the first grid. The bias applied to the grid whose transconductance contours are shown is chosen to be approximately that for which the maximum transconductance is obtained.

The conversion amplification of the pentode mixer may be determined directly from the special curves if the output impedance of the mixer circuit (the impedance Z_L of the i-f transformer) is known. As with the triode mixer, the amplification as a function of the local oscillator voltage may be calculated in terms of the transcon-

[1] Pullen, K. A., Conductance Curves Speed Design of Pentode R-C Amplifiers, *Tele-Tech*, July, 1953.

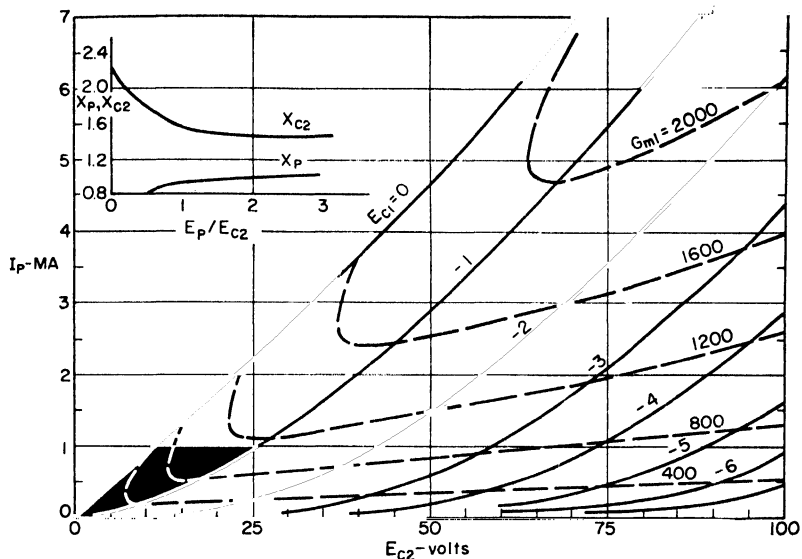



FIG. 42a. Screen characteristics, 6CS6. P_P and P_{c2} = 1 watt.

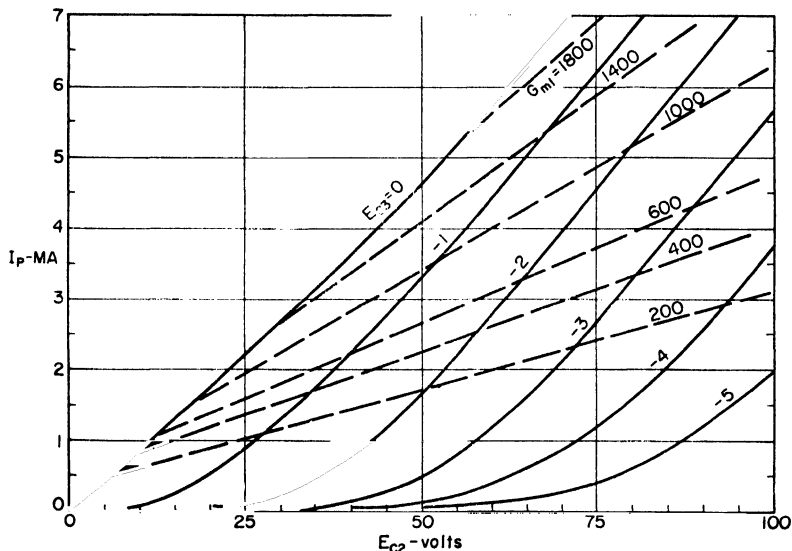



FIG. 42b. Screen-converter curves, 6CS6, E_{c1} = 0.

ductance and converted into terms of K_0 and K_1 in Eq. (42) ($K = -g_m Z_L$). The conversion amplification may then be determined by the help of Table 4.

The curves for the 6CS6 tube in Fig. 42 show several interesting characteristics which are rather typical of conventional mixer tubes. For example, the maximum available transconductance for the first control grid is nearly twice that for the second,

indicating that more effective mixing may be obtained by placing the incoming signal on the first grid. An improvement of the over-all noise figure results since the noise generated is independent of the choice of input grid, whereas the relative output signal is dependent on the maximum transconductance. In addition, the screen current is abnormally high with multicontrol tubes; it may be as much as four times the plate current, reducing severely the available transconductance to the plate and making

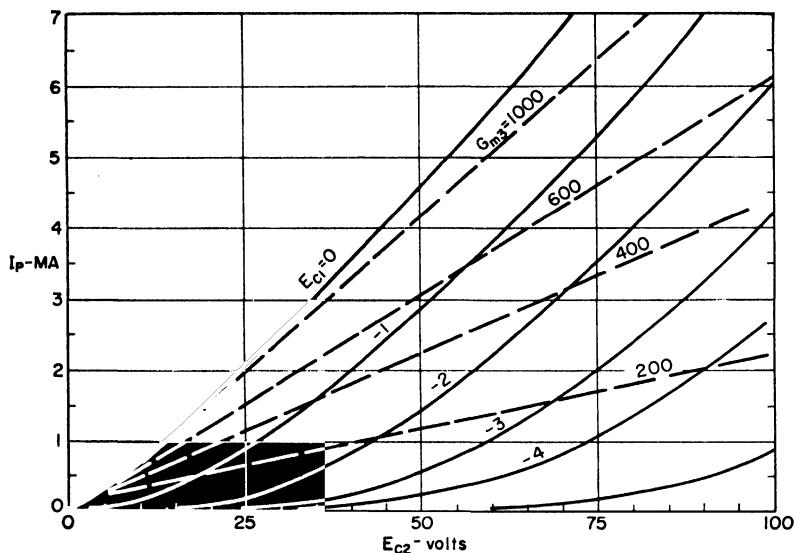



FIG. 42c. Screen-converter curves, 6CS6, $E_{c3} = 0$.

the noise figure even poorer. Furthermore, under the usually recommended operating conditions, a failure of the local oscillator signal results in excessive power dissipation on both the screen and the plate of the tube.

The amplitude of the local oscillator voltage required by the second control grid is appreciably larger than that required by the first grid because of the differences in the transconductances. The peak-to-peak amplitude of the voltage required may be read directly from the converter curves since the local oscillator signal is required to shift the bias sufficiently to change the transconductance from the zero-bias value to a value near zero (at most a quarter of the maximum value). The conversion transconductance available then is approximately a quarter of the difference between the maximum and the minimum values. Usually the screen voltage E_{c2} selected should be sufficiently high to make available as much as possible of the transconductance developed by the tube without permitting damage to the tube with a failure of local oscillator voltage. The value of E_{c2} usually lies between 50 and 75 volts. The final circuit used is shown in Fig. 43.

FIG. 43. Multigrid-mixer circuit.

54. F-m Detectors. Phase and f-m detectors are required to produce an output voltage which is a function of either the applied phase or frequency rather than a

function of the applied signal amplitude. The amplitude of the signal applied to these detectors should be limited to provide constant excitation of the detector; otherwise the output signal is a function of both the frequency and the incoming amplitude of the signal being processed. The detector converts f.m. or p.m. to a.m. and then detects the resulting a.m.; when the amplitude and phase characteristics are linearly variable with frequency or phase, respectively, a.m. proportional to the original frequency or phase is obtained.

If a current of the form indicated in Eq. (45) is applied to a circuit having an amplitude response which varies linearly with frequency, a voltage of the form of Eq. (46) results:

$$i = I_0 \cos (\omega_c t + \theta + m \sin at) \tag{45}$$
$$e = Z_0 I_0 [1 + Sma \cos (at - Pa)] \cos [\omega_c t + m \sin (at - Pa)] \tag{46}$$

where S = slope of impedance characteristic $\Delta Z/\Delta\omega$
P = slope of phase characteristic $\Delta\phi/\Delta\omega$
Z_0 = impedance at carrier frequency
a = angular modulation frequency

Such an impedance is obtained approximately, over a limited frequency range, on the side of the resonance curve of a parallel-tuned circuit or near the resonant frequency of a series-tuned circuit. The resulting voltage is amplitude-modulated at the modulation frequency $a/2\pi$ to a degree Sma. The phase shift of the modulation envelope through an angle Pa and the remaining f.m. or p.m. are usually of no interest. The actual demodulation is made ordinarily in a conventional linear-amplitude detector.

For f-m waves the modulation index is inversely proportional to the modulating frequency $a/2\pi$ and directly proportional to the frequency deviation Δf, where Δf is defined by

$$\Delta f = ma/2\pi = \Delta\omega/2\pi \tag{47}$$

The depth of the amplitude modulation m_a obtained is given by

$$m_a = Sm_f a = S(\Delta\omega) \tag{48}$$

For p-m waves the modulation index (m_p) is independent of the modulating frequency. The output is then distorted unless the demodulated signal is passed subsequently through a circuit whose response is inversely proportional to the modulation frequency.

55. Balanced Frequency Detectors. A single frequency-amplitude converter of any simple type responds to spurious a.m. present in the original signal or introduced by the selective response of the i-f tuned circuits. The differentially combined outputs of two detectors, operated from converters with opposite slopes, produce a net output which is nearly linear with respect to desired frequency deviation but zero at the carrier frequency. This arrangement, when carefully tuned to the carrier, gives an output from spurious a.m. which is proportional only to the instantaneous deviation and is, therefore, masked to a great extent by the desired signal. The response to spurious a.m. may be reduced further by an amplitude limiter ahead of the detector stage, or by using a converter circuit which dynamically maintains the output level substantially constant, regardless of any changes in the input amplitude which are at an audible rate.

The most obvious balanced frequency detector or *discriminator* is a pair of circuits, one tuned slightly above and one slightly below the carrier frequency and provided with separate rectifiers whose outputs are combined differentially. A more satisfactory type is the phase-shift discriminator shown in its more common form in Fig.44*a* and in one of the many possible variations in Fig. 44*b*. It combines two converters of opposite slopes and is operated from a single i-f amplifier or limiter. The primary voltage plus half the secondary voltage is applied to one diode, and the primary voltage plus half the secondary voltage, with reversed polarity, to the other diode.

The shape of the output-frequency characteristic of a phase-shift discriminator is satisfactory when the secondary inductance is equal to or slightly greater than the primary inductance, when the effective Qs of the circuits are approximately equal,

and when these circuits are overcoupled by an amount that depends on the available Q and the required maximum deviation.

The primary, secondary, and output voltages of the phase-shift discriminator of Fig. 44*a* for an input current I at an angular frequency $2\pi(F + \Delta f)$ are approximately

$$E_1 = -2\pi FQLI \frac{1 + j2D}{4D^2 - 1 - K^2 - j4D} \tag{49}$$

$$E_2 = 2\pi FQLI \frac{jK}{4D^2 - 1 - K^2 - j4D} \tag{50}$$

$$\text{Output} = 2\pi FQLI\eta \frac{\sqrt{1 + (2D + 0.5K)^2} - \sqrt{1 + (2D - 0.5K)^2}}{\sqrt{(4D^2 - 1 - K^2)^2 + 16D^2}} \tag{51}$$

where $D = Q(\Delta f)/F$
$K = kQ = Q \times$ coefficient of inductive coupling
η = voltage rectification efficiency of diodes

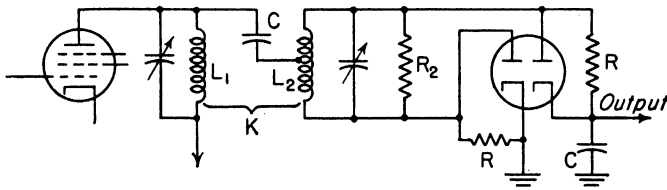
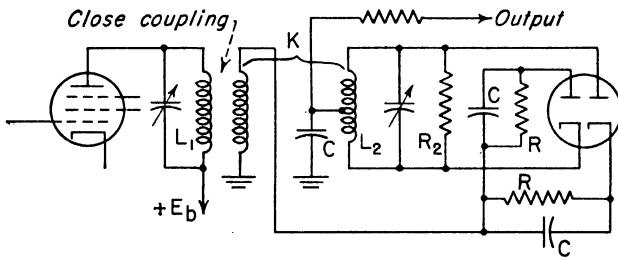
The complex primary and secondary voltages, Eqs. (49) and (50), for $K = 3$ are shown in Fig. 45 for three values of Δf. The phase shifts of the primary and secondary

(a) Conventional

(b) Output balanced to ground

Fig. 44. Phase-shift discriminator.

voltages are in opposite directions as the frequency departs from the center frequency, tending to make the net output more nearly linear with frequency over a wide range.

Universal curves for normalized output, expressed by the radical fraction in Eq. (51), are plotted for $K = 1$, 2, 3, and 4 in Fig. 46. Only one-half of each curve is shown since the output curve for negative deviations is skew-symmetrical with that shown. Actual curves check closely with the calculated curves except when spurious couplings between leads or spurious capacitances to ground modify the results. (The output of the circuit of Fig. 44*b* has the same shape but only half the amplitude.) The value of K to be chosen, not to exceed 3.0, for maximum efficiency is such that $Q(\Delta f)/F$ for maximum deviation is as large as possible while still restricted to a linear portion of the curve. Deviation over too wide a range results in odd-harmonic distortion. Operation with the carrier considerably displaced from the frequency of zero response results in severe even-harmonic distortion.

Both diodes in parallel place a shunt load on the primary while they are in series with respect to the secondary. The resistor R_2 may be added to equalize the circuit Qs. As shown in Fig. 44*a*, the load resistors R are in parallel with the individual diodes, making the net effect of each diode and resistor approximately equal to $R/3$. The capacitors C are for r-f by-pass only.

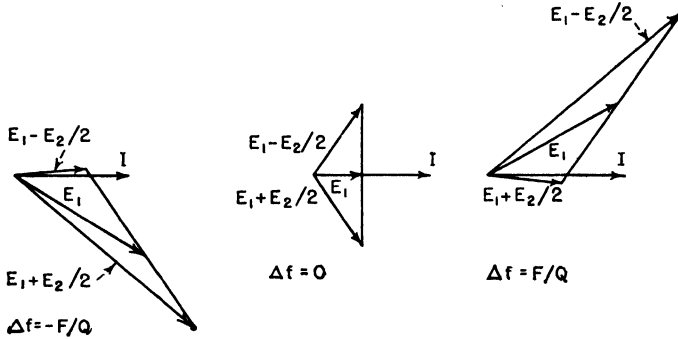



FIG. 45. Voltages in phase-shift discriminator.

A phase-shift discriminator has an advantage over separate tuned circuits in that it may be made linear over greater deviations, since when properly designed the double peaking in the primary circuit and the opposite phase shifts in primary and secondary compensate for some rounding of the maximum response. The tuning is also simpler since the secondary circuit is adjusted to obtain zero response at the center frequency while the primary circuit is adjusted to obtain symmetry of response on either side of center.

FIG. 46. Relative response of phase-shift discriminator.

56. Ratio Detectors. A balanced discriminator very similar to that of Fig. 44*b* but provided with an electrolytic capacitor C_e across the two load resistors in series, as shown in Fig. 47, inherently maintains the output at a level substantially proportional to the frequency deviation and to the average i-f input regardless of a.m. Such a circuit is usually called a *ratio detector* since the ratio of the rectified voltages that appear across the two diodes is approximately equal to the ratio of the open-circuit a-c voltages developed between the two live terminals of the discriminator windings and ground. Many other series arrangements of elements produce equivalent results. Although these circuits are relatively new, they are already in wide use.

The electrolytic capacitor is equivalent to a battery having a voltage equal to the

average rectified voltage at any fixed operating level. Its voltage and the audio output voltage are substantially unaffected by rapid changes in input level, as with a.m., but follow the average input level, as with changes during tuning.

The same rectified current flows through both diodes and the load resistors since their d-c paths are in series. The rectified voltage *across* each diode is equal to its open-circuit peak a-c voltage minus the drop in the equivalent impedance of the source. The rectified current increases during outward a.m., though the division of diode voltages remains the same except for small changes in balance, owing, for example, to small differences between rectification efficiencies of the two diodes. The same is true for the balance for any downward a.m. which does not lower the instantaneous open-circuit peak a-c voltage below that across the electrolytic capacitor.

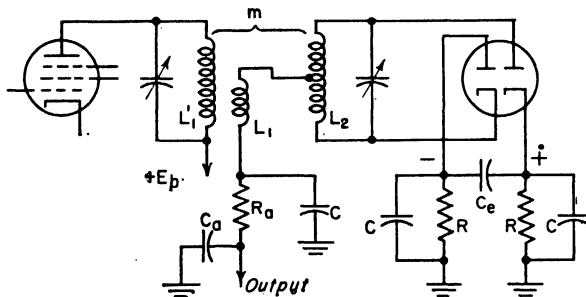



FIG. 47. Ratio detector.

The open-circuit voltages and impedances of the two sources change in the same ratio during frequency deviation, thereby providing differential output as in the conventional discriminator.

Effective nonresponse to a.m. in a ratio detector depends on a correlation of the load resistors with the diode efficiencies and with the impedance of the tuned circuits. This involves proper selections of the self and mutual inductances of the primary and secondary windings and of the r-f by-pass capacitance C. The rectification efficiencies cannot be more than about 50 per cent or the circuit will cut off on downward a.m. and introduce severe distortion. The primary inductance and resistors are usually smaller than in a conventional discriminator. No complete mathematical analysis is known to exist, and designs are checked by tests for linearity of desired output and for residual a.m. at various i-f and modulation levels.

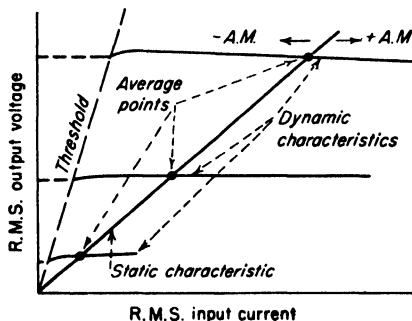



FIG. 48. Dynamic and static characteristic of ratio detector.

Typical characteristics showing net output voltages with variations of input current for a fixed maximum deviation are shown in Fig. 48. The static characteristics are obtained with a slow change of input while the dynamic characteristics are obtained by superimposing a.m. The slope of the dynamic characteristic is proportional to the residual a.m. It may be negative at high input levels; *i.e.*, outward a.m. may cause the net output to decrease. The limiting *threshold*, or degree of downward a.m. to cause diode cutoff, and the ratio of the residual to the original a.m. for about 30 per cent original a.m. are illustrated in Fig. 49.

The residual a.m. is usually divided between balanced and unbalanced effects, as

shown in Fig. 50. In Fig. 50*a* the full line is the deviation characteristic without a.m. and the shaded area covers the limits of residual a.m. for about 50 per cent modulation at the input. The dashed lines show the limits of the envelope in the absence of amplitude reduction. The balanced and unbalanced components are shown (exaggerated) in Fig. 50*b* and *c*. The unbalanced effects may probably be accounted for by small changes in the effective diode input capacitances with level, and by harmonic voltage components introduced in the source impedances by harmonic components in the diode current.

The deemphasis time constant R_aC_a may be made 75 μsec in detectors for broadcast f.m. to compensate for the specified preemphasis.

57. The Gated-beam Detection Circuit. This circuit for the detection of f-m signals has come into common use. It has the interesting characteristics that the variation of the transconductance of the tube used as a function of bias is roughly symmetrical with respect to some value of bias, as is shown in Fig. 39. In addition to the limiting properties resulting from the transconductance symmetry, there exists a negative capacitance and a negative conductance coupling between the input grid and the "quadrature" grid, making possible the development of a quadrature voltage in the latter. The negative capacitance normally is much more effective in introducing coupling at i-f frequencies than is the negative conductance coupling. In the presence of the coupling a tuned circuit can be used to develop a voltage whose phase is a function of the frequency of the incoming signal, thereby producing discriminator action. Limiting is introduced through the action of the symmetrical transconductance characteristic of the first grid. In addition to the limiter curves, Fig. 39 includes a set of negative conductance contours for the quadrature grid. A typical circuit for a limiter-discriminator is shown in Fig. 52.

FIG. 49. Amplitude modulation in ratio-detector output.

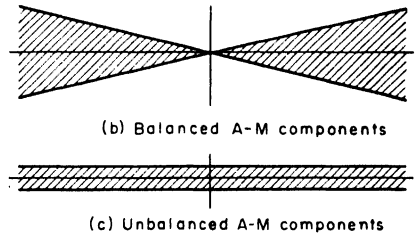
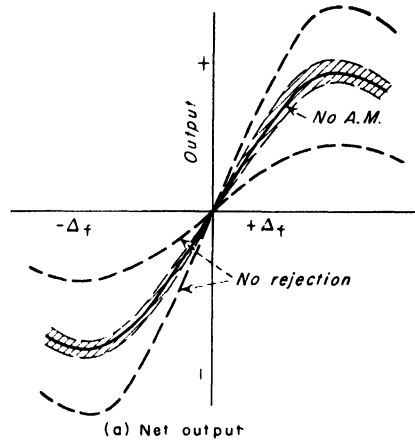



FIG. 50. Residual a.m. in ratio-detector output.

58. Oscillating F-m Detectors. Several types of f-m detector have been developed which utilize the detuning action resulting from coupling of the incoming signal into an oscillator in phase quadrature. Of these, one of the better known is the Bradley detector.[1] It utilizes a multigrid tube having at least two control grids, the first used with the cathode to function as an oscillator, and the second for the input signal. The

[1] Bradley, W. E., Single Stage FM Detector, *Electronics*, **88**, October, 1946.

oscillator is adjusted so that it generates pulses of plate current (class C oscillation conditions). The amplitude of the plate-current pulses is controlled by the input signal, after which the pulses are re-formed to sine waves by a broadly tuned circuit and then coupled back into the oscillator. Such a coupled circuit (Fig. 53) reinserts its voltage into the oscillator itself in quadrature, changing the oscillator frequency and causing the average plate current to vary approximately linearly with frequency.

59. A Filter-type Discriminator. One of the limitations of the conventional discriminators is the rapid loss of output voltage at frequencies somewhat outside the normal operating range. This rapid dropoff, shown at *A-A* in Fig. 51, can cause spurious detection to take place, with the result that tuning of the receiver is made more difficult and also the weak-signal operation is deteriorated. The loss due to the

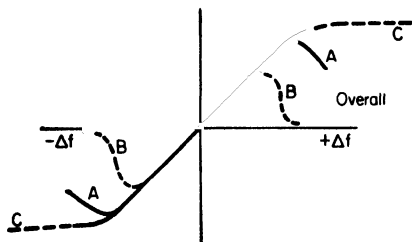



FIG. 51. Discriminator input-output relation.

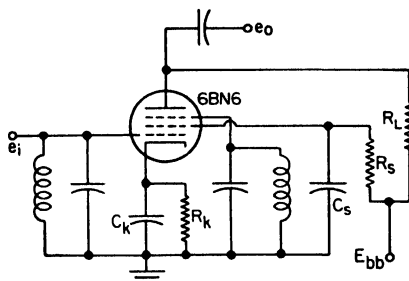



FIG. 52. Gated-beam discriminator.

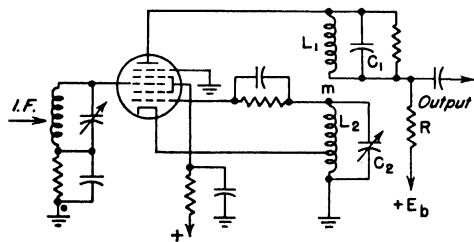



FIG. 53. Bradley f-m detector.

i.f., *B-B*, can be minimized by properly broad-banding the limiter amplifiers ahead of the discriminator. The disadvantage of a sharp loss of output outside the normal pass band may be avoided by the use of a pair of filters, one low-pass, the other high-pass (either *LC* or *RC* filters will serve), to supply the signal voltage to the rectifying diodes. These filters should be designed to have the mid-point of their cutoff range (critical frequency) located on the center frequency of the i-f channel. The voltage output as a function of frequency is shown in *C-C* in Fig. 51. As can be seen, the output voltage outside the detection range is approximately constant, making the discriminator insensitive to frequency excursions outside the detection region. The output characteristics are excellent for AFC use. A control system for AFC which is similar to that described has been reported by Mittelman.[1] A typical circuit is shown in Fig. 54.

60. Phase Detectors. Phase detectors can be used with f-m signals for the detection of modulation if the net phase shift developed in the detector is kept to less than 0.5 radian (otherwise nonlinear operation results). These detectors have recently received

[1] Mittelman, E., *Proceedings of National Electronics Conference*, 1946, p. 503.

a great amount of attention because of the increasing importance of phase-measuring systems in instrumentation, and because of the detection properties obtained with them. Some of the basic properties of these detectors are described in more detail in Sec. 92. Phase detectors may be used to control reactance-modulated oscillators to provide detected signals relatively free of noise in the detected output; a typical circuit for one is shown in Fig. 55.

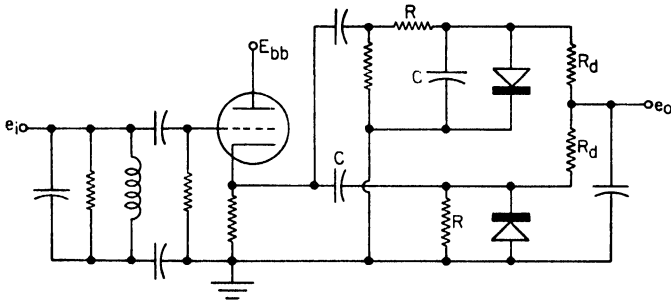



FIG. 54. An *RC* discriminator.

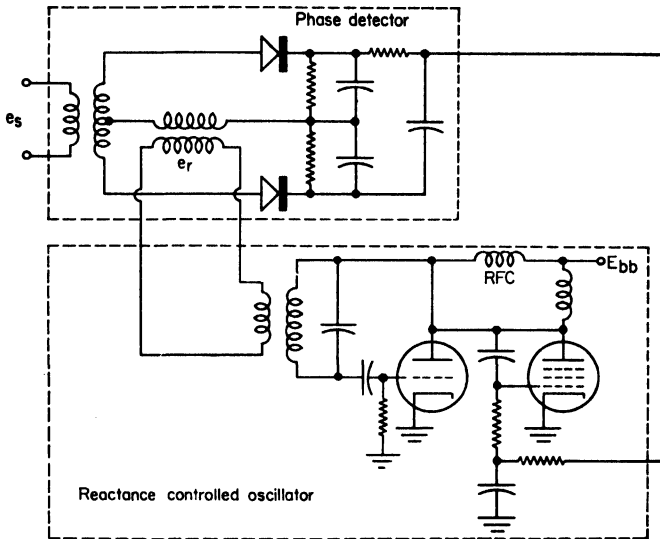



FIG. 55. Phase-lock circuit.

Phase detectors require a local source of signal as well as an external received signal, and they measure the relative phase between the two signals. The output voltage from the phase detector is a function of the cosine of the phase angle between the received voltage and the locally generated voltage, with the result that linear phase detection is available only for a narrow range of phase angles in the neighborhoods of 90 and 270 deg. (This is the reason that phase detectors are seldom used for ordinary f-m communication systems.) Phase detectors with their related control circuits can keep two frequencies in a fixed phase relation to within a few degrees, the exactness being dependent primarily on the phase transients and noise in the original and the reference signals.

Any circuit which may be used for a balanced modulator may also be used as a phase detector, provided only that the active elements are properly matched. Dual diodes are commonly used as active elements because of the relative ease of obtaining adequately matched pairs. A typical circuit using diodes is shown in the portion of Fig. 55 marked Phase Detector.

The phase detector can yield useful data only as long as the phase between the signal and the reference is almost exactly ±90 deg. Since phase detectors are used to lock the frequency of a variable-frequency oscillator to the frequency of a received signal, they must be followed by low-pass circuits in the output which have sufficiently fast response to correct both the phase and the frequency before the phase error actually reaches 90 deg. The band width required in the low-pass filter depends on both the frequency change and the phase sensitivity of the frequency-control circuit.

MODULATORS

Modulators are used to impress intelligence on a signal conveying information from one place to another. Among the commonly used forms are amplitude, phase, frequency, balanced, and pulse modulators. Modulators are very similar to mixers; in fact the principal difference is that, in the modulator, one of the two applied frequencies is the modulation and the other the carrier, whereas in a mixer one frequency is a modulated carrier and the other is a reference carrier.

Modulation may be accomplished on signals of either low or high amplitude. In the former the modulation normally is accomplished early in the process of development of the output signal and is amplified to provide the required final output energy. On high-amplitude signals, the modulation is accomplished in the final amplifier, normally through the use of some form of class C modulator. Each of these arrangements has certain advantages and disadvantages in practical systems; for this reason, each is discussed briefly before the different circuit arrangements available are investigated.

61. Low-level Modulators. Low-level modulators are commonly used in SSB, f-m, and pulse-code systems, but only occasionally for conventional a.m. Unless the amplitude characteristic conveys no information, as in f.m. and sometimes in pulse systems, linear amplifiers must be used with low-level modulators to minimize the amplitude distortion. When properly designed, a system using a low-level a-m modulator followed by linear amplifiers may yield a lower over-all distortion than a high-level system. The introduction of a-m modulation at high power level causes distortion as a result of the apparent reversal of sign of the source impedance of the modulator when the modulation signal voltage is negative (Fig. 56). The resulting distortion can be shown theoretically to be the result of the fact that both the peak and the valley amplitudes of the modulated wave have been reduced without any corresponding reduction of the amplitude of the unmodulated carrier signal.[1] The final amplifier draws power from the modulator when the envelope amplitude is maximum but returns power to the modulator when it is minimum.

FIG. 56. Effect of modulator impedance on amplitude modulator.

62. High-level Modulators. High-level modulators are used almost exclusively in the production of a-m signals and pulse (amplitude) modulated signals. They consist of active devices used in class C amplifiers (control electrode biased statically to cut off all current flow with the carrier removed), the circuit being arranged so that the modulation controls the total carrier energy which can be developed. In the usual

[1] Reinartz, John L., "Increased Audio without Splatter," Eimac Publication.

form, plate modulation, the plate-supply voltage (with triodes) or the plate- and screen supply voltages (with tetrodes and pentodes) are varied by the modulator, causing a variable amount of tuned energy to be available in the plate circuit, producing amplitude modulation of the applied carrier voltage. The system is efficient but is only relatively free of distortion unless extensive feedback networks and artificial loading networks are used to compensate for the nonuniform loading on the modulator.

AMPLITUDE MODULATORS

63. Absorption Modulators. Absorption modulation is obtained by varying a resistance either in series with or in parallel with the load in accordance with some function of the modulating voltage. In Fig. 57*a* suppose that R_v is a resistance which includes the load and which is varied linearly with the modulation or

$$R_v = R_0(1 + M) \tag{52}$$

where M = instantaneous value of the modulation.

The output voltage e_2 is

$$e_2 = \frac{e_1 R_v}{R + R_v} \tag{53}$$

where e_1 = applied voltage
R = resistance of the source

The ratio of e_2/e_1 is plotted against R_v/R in Fig. 58, which indicates that reasonably linear operation is obtained over a small portion of the curve when R_v is small compared to R. The dashed curve is the output voltage across R as a load when R_v is used as a variable series resistor. Physical resistances can be varied over a limited range, say three to one. The usable portion of the curve of Fig. 58 is then limited to the section marked *ab*. The efficiency and the effective degree of linear modulation of the output are low in any case.

Fig. 57. Method of absorption modulation.

The plate resistance of a tube as controlled by the modulating voltage applied to its grid may be used as the variable resistor for absorption modulation in parallel with the output load as shown in Fig. 57*b*. Since the plate resistance is not a linear function of the grid voltage, the over-all linearity may then be improved somewhat by working on a portion of the tube characteristic which tends to cancel the required curvature indicated in Fig. 58. Tests are then made for linearity of load voltage vs. grid voltage.

Fig. 58. Linearity in absorption modulation.

The plate conductance required of the tube is proportional to the resistance R_v when a quarter-wave transmission line or its filter equivalent is interposed between the tube and the load. The required plate conductance g_p is then

$$g_p = \frac{R_v}{Z_0^2} \tag{54}$$

where Z_0 = image impedance of the line.

Absorption modulation has been used to supplement other modulating methods over portions of the modulating cycle.

64. Grid Modulators. Grid modulators operate with carrier and signal voltages applied to the same or separate grids as illustrated in Fig. 59*a* or *b*. The plate current

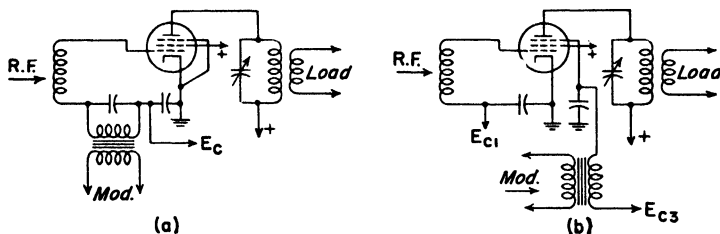



FIG. 59. Types of grid modulators.

may be calculated by Eq. (31) for low levels when the plate current is not swung to cutoff. The action is illustrated in Fig. 60 for the connections of Fig. 59*a* and for a square-law tube. Curve *a* shows the input signal and carrier voltages superimposed, curve *b* the instantaneous plate current, and curve *c* the modulated output voltage with the l-f components filtered out.

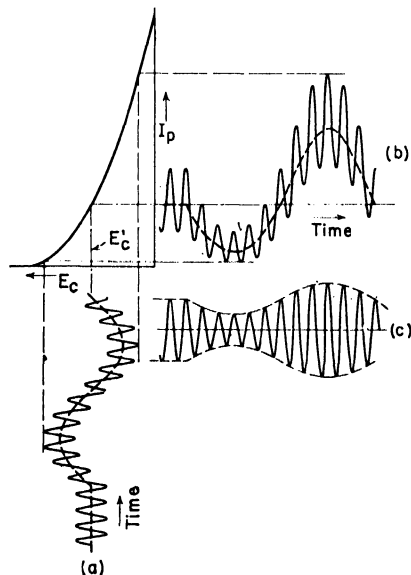



FIG. 60. Low-level grid modulation.

In this case, when the applied voltage about the operating point E_v' is

$$e = E \cos \omega t + S_1 \cos a_1 t + S_2 \cos a_2 t + \cdots \quad (55)$$

where S_1 and S_2 are the signal amplitudes at frequencies $a_1/2\pi$ and $a_2/2\pi$, etc., the useful output current is

$$i = E(A_1 + 2A_2 S_1 \cos a_1 t + 2A_2 S_2 \cos a_2 t + \cdots) \cos \omega t \quad (56)$$

This may be written

$$i = E(A_1 + 2A_2 M) \cos \omega t \quad (57)$$

where M indicates the instantaneous applied modulating signal. The product $M \cos \omega t$, when expanded, produces all the pairs of side frequencies required for the modulated wave. There are no spurious modulation components. However, this mode of operation does not realize fully the power capability of the tube, and the modulation cannot approach unity.

A grid modulator is operated as a class C carrier-frequency amplifier for higher plate efficiency. A tube with a linear grid-plate characteristic is suitable. The bias is adjusted to about twice the value required for plate-current cutoff, and the carrier input voltage is adjusted until the peaks reach halfway between saturation and plate-current cutoff. The superimposed modulating voltage at signal frequency causes the positive peaks to vary widely in value as shown by curve *a* in Fig. 61. Curve *b* shows the instantaneous plate current and curve *c* the useful modulated output voltage with the l-f and carrier-harmonic components filtered out.

Linearity may be tested by observing the output at the test frequency for a range of bias voltages. The exact bias setting is then at the center of the linear portion of the test curve.

Grid modulators have the advantage of requiring small signal input power, particularly when the tubes are not driven to grid current, but have limited ranges of linear modulation and plate efficiencies of only 20 to 30 per cent. They are used ordinarily at low power levels.

Grid modulators are used for television signals since it is difficult to obtain reasonable operation with high-level plate modulation over the required wide band of television modulation frequencies. Triodes may be used in grid modulators if neutralized to prevent h-f feedback. The voltage, current, and power in the plate circuit have the following approximate relations for sine-wave modulation:

$$
\begin{aligned}
E_{\max} &= E_c(1 + M) \\
W_{\max} &= W_c(1 + M)^2 \\
i_{p\max} &= I_c(1 + M) \\
W_{av} &= W_c\left(1 + \frac{m^2}{2}\right)
\end{aligned} \tag{58}
$$

where the subscript c indicates the conditions for the carrier alone.

Fig. 61. Class C grid modulation.

An extra tube which is made to operate by adjustment of its grid and plate potentials over a portion of the modulation cycle to supplement the first modulator tube is sometimes used to obtain higher efficiency and at the same time preserve or even improve linearity of modulation.[1]

65. Plate Modulators. The constant-current plate modulator utilizes an a-f choke coil in the circuit which supplies plate power to r-f and a-f amplifier tubes, as shown in Fig. 62. The total plate current remains constant by virtue of the inductance of the choke. The instantaneous audio plate voltage is added to the plate supply voltage and over the audio cycle changes the latter to $E_b(1 + M)$. The r-f inductance L prevents the loss of h-f power in the audio tube, and C prevents the short circuit of the audio plate voltage.

The a-f tube, frequently called the *modulator*, supplies modulating power, but the actual modulation occurs in the plate circuit of the r-f tube.

The plate current of the a-f tube cannot be reduced to zero during the modulation cycle without introducing audio distortion. It is necessary, therefore, to operate the

[1] Terman, F. E., and J. R. Woodyard, A High-efficiency Grid-modulated Amplifier, *Proc. IRE*, **26**, 929–945, August, 1938.

audio tube with a higher zero-signal plate current than the radio tube to reach unity modulation without audio distortion. This is done by applying a higher plate voltage to the audio tube, either by using a by-passed resistance in the plate circuit of the r-f tube or by supplying the audio tube (or tubes) through transformer coupling as illustrated in Fig. 63. A further improvement is indicated in Fig. 63, since saturation of the transformer core is prevented by eliminating the d-c magnetizing component and since even-order audio harmonics are canceled by the push-pull arrangement. The efficiency of the system is increased by operating the audio tubes in push-pull class B. In transformer coupling the total d.c. is no longer constant but varies with the modulation.

Fig. 62. Circuit for plate modulation.

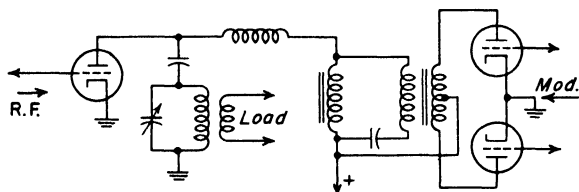



Fig. 63. Transformer-coupled plate modulator.

The voltage, current, and power in the plate circuit have the following relations for sine-wave modulation:

Tube and circuit voltage,

$$E_{\max} = (1 + m)E_b$$

R-f input power,

$$W_{\max} = (1 + m)^2 E_b I_b$$

Average r-f input power,

$$W_{av} = \left(1 + \frac{m^2}{2}\right) E_b I_b$$

Average output power,

$$W_0 = \eta \left(1 + \frac{m^2}{2}\right) E_b I_b$$

Average audio input power,

$$W_a = \frac{m^2}{2} E_b I_b$$

R-f plate loss,

$$W_p = (1 - \eta)\left(1 + \frac{m^2}{2}\right) E_b I_b \tag{59}$$

where E_b and I_b = d-c supply voltage and current
m = degree of modulation
η = plate efficiency of radio tube

The plate efficiency is high (often 0.7 to 0.8), and the chief disadvantage of plate modulation is the large amount of audio power which must be supplied.

The radio tube is operated as a class C amplifier with bias at approximately two times cutoff. Since the plate losses are 50 per cent higher with full modulation than for unmodulated output, the tube must be used at two-thirds its rated power. Low-mu triodes are suitable and ensure low plate and grid voltages.

Grid-leak bias helps in obtaining linearity up to complete modulation. Linearity may be checked by direct adjustment or by test at 60 cycles with a proper plate load for a range of plate voltage from 0 to $2E_b$.

66. Modulated Oscillator. Plate modulation was originally applied directly to the oscillator tube and circuit. Practically full modulation may be obtained with excellent linearity, but the arrangement has the disadvantage of introducing f.m. The frequency of the oscillator varies with the plate voltage, and, since in plate modulation this varies between $2E_b$ and 0 during full modulation, the oscillator frequency deviates from its mean value with the modulating signal.

The same circuits are used between the two tubes, and the same voltage, current, and power relations hold as with a plate-modulated amplifier. Linearity is obtained by adjusting the value of the grid leak.

Modulated oscillators are now considered suitable only for test equipment in which the f.m. is not objectionable.

67. Semiconductor Modulators. Copper-oxide and other forms of semiconductor rectifiers are applicable in bridge modulators and are used widely in carrier-current telephony. They function as carrier-operated switches for opening, shorting, or reversing the elements carrying the modulating currents. For such purposes they are the familiar dry rectifiers for 60-cycle power in miniature. Each unit consists of a small copper disk, one side of which is oxidized and placed in contact with a soft metal. Electrons flow most easily from the copper to the soft metal. The resistance of a $\frac{3}{16}$-in. unit at low frequencies is of the order of 100 ohms in the forward direction and from 10^4 to 10^5 ohms in the reverse direction. The current-carrying capacity is roughly proportional to the area of the disk, which may be $\frac{1}{16}$ in. or more in diameter. The voltage rating is increased by placing several disks in series. Bridge units are usually hermetically sealed with only the terminals brought out.

Copper-oxide rectifiers are not suitable for use at frequencies much above 1 Mc, except at low impedance levels on account of inherently large shunt capacitance. They are compact in size and eliminate the heater connections necessary in similar circuits using vacuum tubes. They maintain a satisfactory balance in carrier-suppression circuits using balanced modulators.

The power-series current equation for a copper-oxide unit converges slowly, and its characteristics are expressed more easily quantitatively in terms of resistance for different applied voltages. For voltages in the reverse direction and for less than 0.02 volt in the forward direction, the resistance is high and substantially constant. For forward voltages between 0.02 and 0.6 volt, the resistance is approximately

$$r = r_0 \epsilon^{-ke} \tag{60}$$

where k is a constant which may be as great as 18. For forward voltages larger than 0.6 volt, the resistance is low and nearly constant.

Representative circuits using copper-oxide units are shown in Figs. 64 and 65. In these figures, f_c and f_s indicate voltage sources at the carrier and signal frequencies, respectively. The impedances of the carrier source, signal source, and load are Z_c, Z_s, and Z_L, respectively. The forward-conducting direction of the units is shown by the arrows.

FIG. 64. Bridge modulator using nontube rectifiers.

Since the carrier and signal voltages are applied to conjugate terminals of the rectifier bridges, the resistance effects are balanced. In Fig. 64 the output signal is short-circuited for one polarity of the carrier cycle. In Fig. 65, the effective connections between the signal source and the load are reversed as the polarity of the carrier changes. This arrangement is called a *double-balanced* or *ring modulator*.

The frequencies of the current components produced in the individual rectifier units are determined qualitatively by an expansion of each of the terms of a power-series equation for the current. Current components of frequencies equal to the sums and

differences of the integral multiples of the carrier and signal frequencies appear in each unit in the forward direction. These combine additively or differentially in the connected circuits depending on the polarity. In flowing through the circuit impedance these current components produce voltages of the same frequency which are reimpressed upon the rectifier units. The final result may be obtained quantitatively only by a series of approximations. The frequencies of the components appearing in the circuit impedances are indicated in Figs. 64 and 65, where n is any whole number or zero, n_o is any odd number, and n_e is any even number or zero.

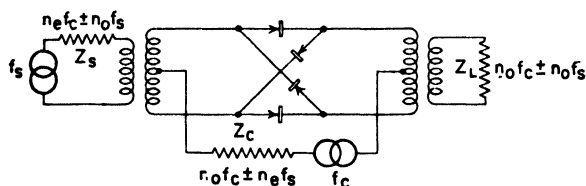



Fig. 65. Double-balanced or ring modulator.

The output impedances are designed as filters to eliminate voltages at frequencies involving undesired multiples of f_c. The useful output is at a frequency $f_c \pm f_s$, and, when double modulation is used, the carrier frequency f_c is eliminated.

By making the carrier voltage large in comparison with the signal voltage, the terms involving multiples of f_s may be reduced satisfactorily in magnitude. The units are operated with about 0.5 volt carrier across each disk in the forward direction. The optimum impedance in the signal and output circuits is

$$Z = \sqrt{R_f R_r} \tag{61}$$

where R_f and R_r are the resistances in the forward and reverse directions, respectively. The loss in conversion is then only 6 to 8 db.

Copper-oxide bridge modulators differ from van der Bijl vacuum-tube balanced modulators in that they transmit in either direction. They function equally well as modulators or demodulators.

The same circuits may be used for solid-state crystal diodes as for copper-oxide rectifiers. They have low self-capacitance, and single units are operable to frequencies of several hundred megacycles.

68. Balanced Modulators. When carrier voltage is applied in phase and modulating voltage is applied in push-pull to the grids of two modulator tubes, the carrier is balanced out in a push-pull output load. The circuit shown in Fig. 66 with two neutralized triodes is typical.

Fig. 66. Balanced-modulator circuit.

For exact balance of tubes and transformers, and over the range of signals for which the modulation characteristic is linear, the useful modulated output current from the two plates is

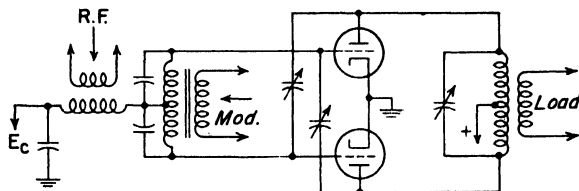
$$i_1 = E_c(K_0 + K_1 e_m) \cos \omega t \tag{62}$$

and

$$i_2 = E_c(K_0 - K_1 e_m) \cos \omega t \tag{63}$$

where E_c = maximum carrier voltage
e_m = instantaneous modulating voltage

The effective input current to the tank circuit is

$$i = 2K_1 E_c e_m \cos \omega t \tag{64}$$

which contains only the side bands. For a single modulating frequency $a/2\pi$, this reduces to

$$i = K_1 E_c e_m [\cos (\omega + a)t + \cos (\omega - a)t] \tag{65}$$

The voltage developed in the output circuit is

$$e = K_1 E_c e_m Z [\cos (\omega + a)t + \cos (\omega - a)t] \tag{66}$$

where Z = load transfer impedance.

This arrangement is used in suppressed-carrier transmission systems. It has the advantage of balancing out any even-harmonic distortion due to departure of the modulation characteristic from linearity, not considered in the above equations.

When the modulating input voltages to the two grids, the tube coefficients, and the effective load transfer impedances for the two tubes are unequal, the net tank circuit voltage is

$$e = E_c[K_0'Z' - K_0''Z'' + (K_1'e_m'Z' + K_1''e_m''Z'') \cos at] \cos \omega t \tag{67}$$

where the $'$ and $''$ values are for the first and second tubes, respectively. Some of the carrier remains when the balance is not perfect.

69. Applications of Balanced Modulators. Any combination of active devices may be used in a balanced modulator if the circuit is so arranged that it is symmetrical with respect to one of the input signals, causes one signal to be balanced out statically in the output circuit, and has sufficient nonlinearity so that neither signal is balanced out in the presence of a signal in the symmetrical input. This balancing out of the static carrier results in the development, in the modulation, of a complete set of modulation side bands but no carrier.

The carrier component can be reinserted in the output obtained from the balanced modulator; if it is introduced in phase with the original carrier signal, then either conventional a.m., controlled-carrier a.m., augmented-carrier a.m., or any of the related forms of a.m. may be obtained. If the carrier is reinserted in quadrature, however, phase modulation is generated; the phase deviation permitted is less than ½ radian if linear modulation is required. If the carrier is not reinserted, and one of the side bands is suppressed, then a single-side-band suppressed carrier (SSSC) is obtained.

The suppression of one side band may be obtained through the use of filters or through the use of balancing circuits. The so-called phase system of developing SSSC signals uses the phasing or balancing-circuit technique to eliminate one side band; it uses two balanced modulators along with two carrier voltages which are in phase quadrature along with modulation signals which also are in phase quadrature. Combining the output of the two modulators gives the desired SSSC signal.

A filter or rejection circuit also may be used to eliminate one of the side bands. For SSSC service, crystal filters are usually used, and for straight suppressed side band (but not necessarily suppressed carrier, as in TV) a constant-impedance band-rejection filter may be used. Crystal-lattice filters are used extensively for the rejection.

PHASE MODULATORS

Phase modulators are devices which apply modulation to a carrier signal by advancing and retarding the phase of the carrier voltage in step with the modulation information. Three common methods for the generation of p-m waves are (1) use of the balanced modulator with the carrier reinserted in quadrature, (2) the phasitron, and (3) the use of the *Serrasoid modulator* (Sec. 72).

70. Balanced Modulator Method of P.M. When carrier voltage is added into the output of a balanced modulator in such a way that the carrier phase lags or leads by 90 deg the phase of the phasor sum of the modulation side bands, then the side bands in effect retard or advance the phase of the carrier, giving p.m. and a small amount of a.m. (Fig. 67). The amount of modulation which can be introduced

linearly with such a modulator is limited to less than 25 deg, the maximum angle at which the sine can be approximated by the radian value of the angle. The resultant phase-modulated voltage e_p varies in phase from the new carrier E_c' by the angle

$$\phi = \tan^{-1} \frac{E_s \cos at}{E_c'} \tag{68}$$

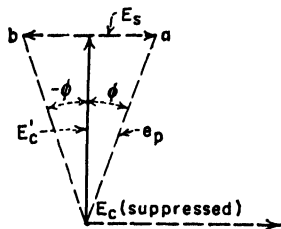
which, for angles less than about 25 deg, is approximately

$$\phi \cong \frac{E_s \cos at}{E_c'} = m_p \cos at \tag{69}$$

FIG. 67. Vector relations in p.m.

The resultant voltage varies only slightly in magnitude and is

$$e_p = E_c' \cos (\omega_c t + m_p \cos at) \tag{70}$$

When modulating signals at more than one frequency are present, the coefficients m_1, m_2, etc., are proportional to the original a.m., and p-m waves are produced.

The small phase departure of less than 25 deg may be increased by frequency multiplication of the instantaneous frequency. The new voltage is then

$$e_p' = E_c' \cos (n\omega_c t + mn \cos at) \tag{71}$$

71. Phasitron. A tube, called the *phasitron*, gives a large amount of phase modulation proportional to the current in a coaxial modulating coil surrounding it.[1] The tube utilizes a radial electron stream formed into a disk with a wavelike edge. The scalloped beam is made to progress at a constant average angular velocity about the cathode by deflecting electrodes excited by a crystal oscillator, and alternately to strike anode areas connected to the opposite terminals of a balanced output circuit. The phase of the scalloped edge is advanced or retarded by the modulating field.

When modulating voltages of constant amplitude are applied to the coil, the current is inversely proportional to their frequency and f.m. is obtained directly. The phase shift over which the modulation is linear may be more than 360 deg so that less frequency multiplication is required to obtain the standard system deviation for f-m broadcasting than with conventional phase modulators.

72. The Serrasoid Modulator. The serrasoid modulator uses a crystal-controlled saw-tooth generator in conjunction with a low-frequency modulation voltage to produce a variable triggering time for exciting a blocking oscillator (Fig. 68). When the rising portion of the saw tooth crosses the control-voltage level, the blocking oscillator produces a pulse which can be filtered to provide a p-m carrier at a small multiple of the crystal oscillator frequency.

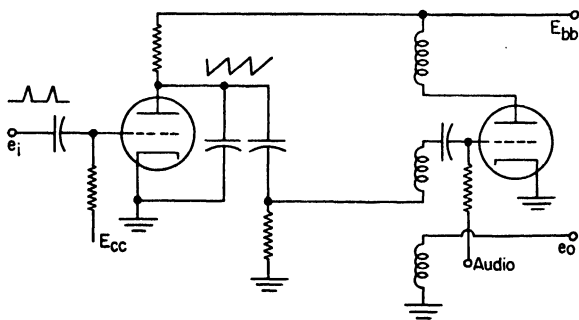



FIG. 68. Serrasoid modulator.

The excitation pulses are shown at e_i; they are filtered by a tuned circuit to reduce them

[1] Adler, R., A New System of Frequency Modulation, *Proc. IRE*, January, 1947, p. 25.

to a variable-phase sine wave. The resultant signal may then be multiplied as required.[1]

FREQUENCY MODULATORS

Frequency-modulation circuits are closely related to phase-modulation circuits since frequency is actually the rate of change of phase. If, therefore, a modulation signal is first integrated and then phase-modulated, the signal produced is identical with that obtained by direct frequency modulation. As a result, phase modulators are usually used to produce f.m. because of the better frequency stability which can be obtained with them.

Frequency-modulated waves are obtained by the method described in Sec. 70 when the modulating signal is passed through a filter whose response is inversely proportional to the signal frequency. The instantaneous frequency is multiplied several hundred times before the output is applied to the antenna.

73. Reactance-tube Modulator. Frequency modulation may be obtained, however, by the use of an oscillator coupled to a reactance tube, which introduces a fictitious capacitance, dependent on the modulation, into the oscillator circuit. This capacitance is made variable by having the modulation voltage vary the transconductance of the reactance tube, thereby causing the introduction of a variable magnitude of quadrature current into the oscillator (Fig. 69). The quadrature current appears to change the reactance of the tuned circuit of the oscillator, varying the frequency in step with the modulation voltage. The linearity of the resulting modulation is critically dependent on the characteristics of both the circuits and the reactance tube itself.

FIG. 69. Reactance-tube f-m modulator.

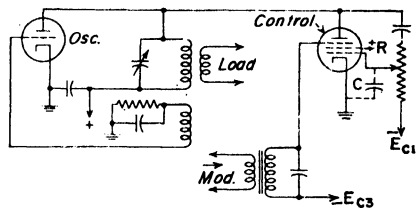
Either an integrator or a differentiator circuit may be used to couple the oscillator signal into the grid of the reactance tube; then the quadrature current from the reactance tube is reinserted into the tuned circuit of the oscillator to vary either the inductance or the capacitance. An integrator followed by an amplifier generates a capacitive current, and a differentiator circuit produces an inductive current. The magnitude of the quadrature current is controlled by the magnitude of the modulation voltage and by the transconductance of the reactance tube for each value of the modulation voltage.

A typical circuit is shown in Fig. 69 as one of many possible arrangements. Radio-frequency voltage, shifted in phase 90 deg from that appearing across the tank circuit, by means of resistance R in series with capacitance C, is applied to grid 1 of the control tube. The plate current of the control tube is substantially 90 deg out of phase with the tank voltage and provides an effective reactance which is controlled in magnitude by the modulating voltage on grid 3.

The change in oscillator frequency is proportional to the instantaneous modulating voltage when the control tube operates on a linear part of its characteristic and when the total change in effective reactance is small compared to the net average reactance.

Any component of control-tube plate current not at 90 deg with the tank voltage will introduce a.m. This is eliminated by adjusting the phase shift to grid 1 of the control tube.

The curve of frequency deviation from the carrier frequency should be linear with respect to the amplitude of the modulating voltage. It may be checked by applying direct voltages, over the operating range, to the modulating grid and observing the oscillator frequency.

The circuit illustrated in Fig. 69 provides f-m waves when the modulating voltages

[1] Day, J. R., Serrasoid FM Modulator, *Electronics*, October, 1948, p. 72.

are proportional to the amplitude of the signal. If the signal is passed through a filter whose output is proportional to the signal frequency before application to the control tube, p-m waves are produced.

PULSE SHAPING AND FORMING

Pulse shaping and forming may be accomplished either by the use of clamp or limiter circuits in connection with differentiator or integrator circuits, or through the use of regenerative pulse-forming circuits. The former are used when the duration of the shaped pulses is relatively unimportant, and the latter when the duration is important.

The shaping and forming of pulses with nonregenerative circuits requires combinations of amplification, clamping, and differentiation. The incoming wave should have characteristics resembling a pulse or a square wave if the stability of initiation is important. For the time jitter to be small, the initiating signal must switch rapidly across the gap between the lower and upper clamp levels. Consequently, rapid switching and clean clamping or limiting are required in conjunction with broadband amplification and differentiation circuits to form sharp stable pulses.

74. Active Pulse-forming Circuits. Active or regenerative pulse-forming circuits are used to generate pulses for devices like radar sets or digital computers; they are usually triggered by the signal from a nonregenerative pulse-forming circuit. Among some of the more common regenerative pulse-forming methods are:

1. One-shot multivibrator
2. Reflection on delay lines
3. Blocking oscillators
4. Delay lines in negative-feedback circuits
5. Combinations of circuits

Multivibrators and nonregenerative circuits can produce pulses having durations greater than 2 μsec, whereas the remaining circuits in the above list normally are used to generate pulses having durations between about 0.1 and 100 μsec.

75. The Single-shot Multivibrator. Normally the single-shot MV is used to convert trigger pulses into pulses having controlled amplitude, duration, and wave shape. The initiation pulse required to initiate any form of regenerative switch, and particularly multivibrators, must have both a greater amplitude and a longer duration than expected since the pulse duration must be sufficient to increase, dynamically, the loop gain of the regenerative device to a value greater than unity. The effects of capacitive loading make a sizable reduction in the apparent trigger amplitude, often with the result that the initiation pulse gets the device only partially triggered; switching, if it occurs at all, occurs only after a considerable delay. Delay times of between 10 and 30 μsec are by no means unusual.

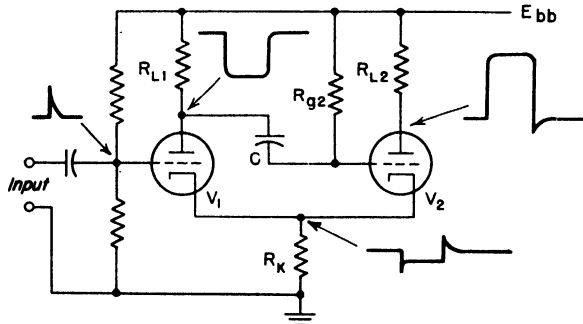



FIG. 70. Cathode-coupled one-shot MV.

One of several possible types of one-shot MV is the cathode-coupled circuit shown in Fig. 70 together with typical output wave forms.[1] It may be adjusted for pulses of 2 to 2,000 μsec duration by changing the capacitance C and/or the various resistors. Positive or negative pulses are available at the plates of V_2 and V_1, respectively. A negative trigger may be used at the grid of V_2 instead of the positive trigger shown at the grid of V_1. In any case the dura-

[1] M.I.T. Staff, "Principles of Radar," Chap. 2, Secs. 13 to 16, McGraw-Hill, 1946.

tion of the trigger should be less than the desired pulse, and no other extraneous pulses should be coupled to either grid.

V_2 conducts between pulses since its grid-cathode voltage is approximately zero, while V_1 is cut off with a sufficiently high cathode voltage. A negative trigger at the grid of V_2 reduces the current and the cathode voltage to a point where V_1 starts to draw current. Almost instantly the negative voltage developed in the plate load of V_1 drives V_2 to cutoff. The current in V_1 is then temporarily stable as determined by the tube characteristic, resistors R_{L1} and R_k, and the terminal voltages. The negative peak voltage at the grid of V_2 is substantially equal to the drop across R_{L1} although it is modified slightly by the interelectrode and circuit capacitances. This grid voltage rises exponentially toward E_{bb} until V_2 again draws current. A positive voltage is regeneratively coupled almost instantly to the grid of V_2 via increased cathode voltage and reduced drop across R_{L1} so that V_2 momentarily draws more than normal current while V_1 is cut off. Grid current in V_2 then quickly reduces the grid voltage to normal, and the original conditions are restored.

The duration of the pulse depends on, but is less than, the time constant of C in series with R_{g2} and the combination of R_{L1} and the plate resistance of V_1. The pulse may be lengthened by returning the grid resistor R_{g2} to cathode rather than to E_{bb} since the voltage across the charging circuit for C during the pulse is then reduced. The duration of the pulse may be controlled to a limited extent by changing the size of R_{L1} since it determines the peak voltage to which C is charged during the pulse. Likewise this peak may be controlled by adjusting the tap of the voltage divider for the bias voltage applied to the grid of V_1. The pulse may be lengthened by increasing C, R_{L1}, or R_{g2}, or by reducing the voltage at the grid of V_1.

76. Pulse-forming Networks. Low-loss lines with distributed constants, open or shorted at the far end, or their network approximations, are used at various voltage levels for forming rectangular pulses of duration equal to the round-trip delay therein.

Figure 71*a* shows an open-end line with several sections of L and C approximating a uniform line. Figure 71*b* is a network simulating the sectionalized line more suited for operation at high voltage since, when completely charged, all the energy is stored in the single capacitance C_{st}, which is equal to the sum of the Cs in Fig. 71*a*. The circuits L_1C_1 and L_2C_2 have natural frequencies equal to $1/t_w$ and $2/t_w$, respectively. C_{st} is also equal to $t_w/2R_c$, where R_c is the characteristic impedance of the line. A pulse of half the initial voltage on the charged line appears across the load during the time the wave front travels to the open end and returns, when the line is discharged through a resistance equal to the characteristic impedance R_c. The traveling voltage pulse is negative and is reflected from the open end of the line with the *same* polarity to cancel the remaining voltage at the load. The discharge is by smaller proportional steps when the load resistance is greater than R_c, by decreasing alternately reversed steps when smaller.

(a) Sectionalized

(b) Simulated

(c) Circuit

(d) Pulse

FIG. 71. Discharge of open-end delay network.

Figure 72 shows a circuit in which a thyratron V_2 is made to act like a low-resistance electronic switch to discharge an open-end line by a positive trigger pulse on its grid. The network is desirably discharged through a load matched to the characteristic impedance. The by-pass diode V_3 provides current during any overshoot of voltage due to load mis-

match so that the plate of V_2 cannot become negative. The network recharges relatively slowly from the source E_{bb} through the charging diode V_1 and the inductance L after the pulse. This must be after the thyratron has deionized so that it will not short-circuit the source. The voltage on the line increases as the half cycle of an oscillatory charge to almost $2E_{bb}$, where it remains owing to the switching action of the diode V_1 until again triggered.

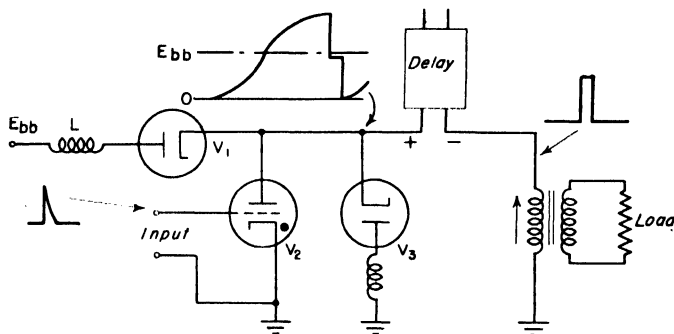



FIG. 72. Resonant charging of pulse-forming delay network.

The charging diode and the inductance may be replaced by a high resistance at the expense of charging the line to only the voltage E_{bb}. A rotary-gap switch is sometimes used instead of the thyratron; also the charging may be from an a-c source.

Figure 73*a* shows a shorted-end line with sectionalized L and C. In this case t_w is $2L_{st}/R_c$. Figure 73*b* is a network which simulates the uniform line on short circuit. When the line is connected to a source of constant current by opening a switch as in Fig. 73*c*, a pulse of voltage equal to $R_cI/2$ appears until the initial wave travels to the end of the line and returns with *reversed* polarity to cancel the initial voltage, as shown in Fig. 73*d*.

Figure 74 shows a circuit in which such a shorted-end line may be used to develop a short pulse from a longer input pulse. When V_1 is suddenly made conducting by a wide input pulse, a negative voltage pulse appears at its plate for the duration of the round-trip delay of the line. The reverse occurs as the initial pulse ends. Tubes V_2 and V_3 may be used to clip the pulse by plate current cutoff so that only a rectangular pulse of the desired duration appears in the output. The approximate wave forms at each grid are shown. Tube V_3 acts as a cathode follower with low internal impedance and, therefore, has good regulation during the pulse.

(a) Sectionalized

(b) Simulated

(c) Circuit

(d) Pulse

FIG. 73. Pulse form in shorted-end delay network.

77. The Blocking Oscillator. The blocking oscillator is a special form of sine-wave, or inductively coupled, oscillator, one in which the feedback is so large during the major portion of the conducting cycle that a large amount of grid current is drawn, damping the sine wave and storing sufficient charge in the grid capacitor to ex-

tinguish conduction after a single distorted cycle has been developed. Initiation of the blocking oscillator is dependent on the rate of change of voltage in the initiating pulse, with the result that accurate triggering is even more dependent on pulse shape than it is in an MV.

A typical circuit for a blocking oscillator is shown in Fig. 75. Proper operation is critically dependent on the existence of a source of stored energy, usually the capacitor C_s shown in Fig. 75, connected from the cathode of the tube to the supply end of the plate winding of the transformer. Either absence of the capacitor or the use of a smaller capacitor than is required causes erratic operation.

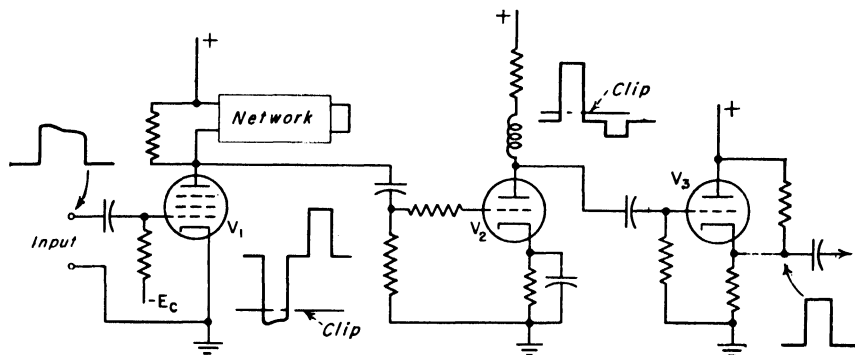



FIG. 74. Pulse forming by shorted-end delay network.

The blocking oscillator develops its pulse through damping and amplifier saturation. Either the first half cycle or the first positive overshoot of the plate voltage due to transformer resonance may be used directly with the blocking oscillator to modify the grid or plate supply voltage as required. The pulse shape is then somewhat rounded, and better results are obtained after clipping in a driving tube. An iron-core pulse transformer couples the plate to the grid in proper polarity for oscillation, which is prevented between pulses by a large negative grid voltage. The output load is coupled from the same transformer. The plate current is abruptly increased by regenerative action when the grid voltage is made sufficiently positive, thereby initiating a half cycle of violent oscillation. Bias voltage is produced thereby to prevent further regenerative operation until again triggered.

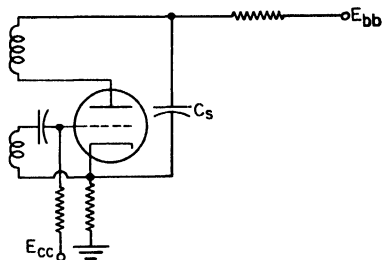



FIG. 75. Blocking oscillator.

The voltage on C_2 in Fig. 76 is increased at each trigger pulse through C_1 and diode V_1 by an amount equal to $C_1/(C_1 + C_2)$ times the difference in voltage of the pulse and the voltage which is on C_2 immediately preceding the pulse. The tap for the cathode of blocking oscillator V_3 may be adjusted so that one or more input pulse steps are required to raise the grid potential above plate-current cutoff, thereby providing operation at the trigger repetition frequency, or some submultiple thereof. This expedient is useful where *division* of the repetition rate is desired, as, for example, when the average load in the equipment corresponding to a low repetition rate must not be exceeded. The first half cycle of a violently regenerative oscillation is started when plate current flows. Grid current is then drawn through diodes V_1 and V_2 in series, discharging C_2 so that the oscillation is not repeated until triggered again.

The grid current during each initial half cycle of the blocking oscillator of Fig. 77 charges C_1 to such a voltage that the tube then remains cut off for a considerable interval, depending on the time required to discharge C_1 through R_1 and R_2 in parallel. This determines the repetition rate and no trigger pulse is needed. Plate current is terminated at the end of

each half cycle, and the first positive overshoot of plate voltage due to transformer resonance provides the useful output. This allows considerable storage of energy $(LI^2/2)$ in the field of the transformer, which develops a voltage somewhat less than $I\sqrt{L/C}$ during the first damped half cycle after plate-current cutoff. The pulse duration is determined by the transformer inductance and circuit capacitance.

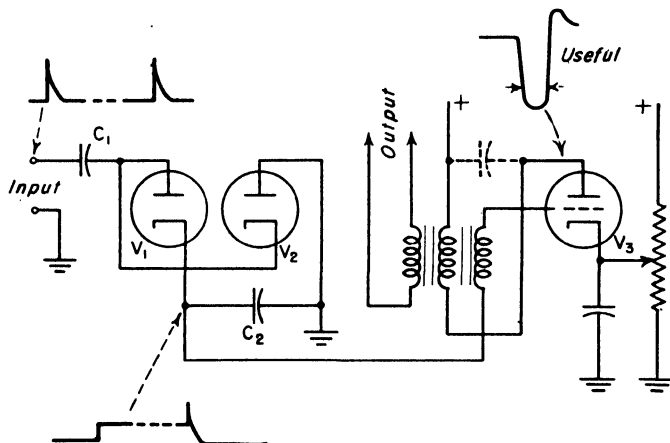



FIG. 76. Triggered blocking oscillator.

78. Termination of Pulse in Oscillator Circuit. The circuit for the supply of grid or plate voltage to triode r-f oscillators may be so arranged that the duration of the pulse, when once triggered, is determined in the oscillator circuit itself. Such arrangements are used at low power and where the pulse shape is not critical.

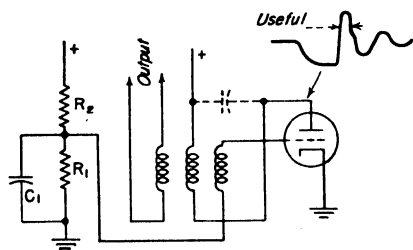



FIG. 77. Self-pulsed blocking oscillator.

Figure 78 shows a circuit for a triode oscillator in which the pulse length is determined by a combination of the time constants of the grid circuit of the driving tube V_1 and of the cathode circuit of the oscillator V_2. The cathode of V_2 is maintained at a sufficiently high potential between pulses by the bleeder to prevent the flow of plate current. The leading edge of the input pulse raises the oscillator grid voltage and initiates oscillation. C_2 then charges slowly to raise the cathode voltage while the grid supply voltage falls as the charge

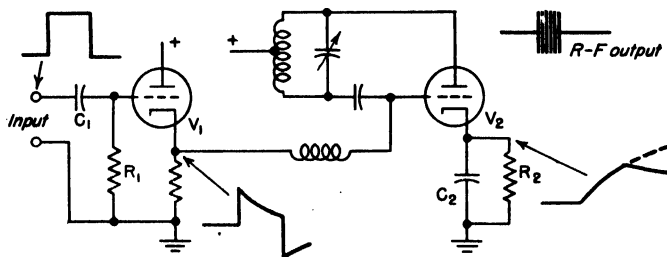



FIG. 78. Time constants of supply circuits determine duration.

on C_1 leaks off through R_1. After an interval which may be adjusted by changing C_1, R_1, or C_2, the tube can no longer oscillate and the pulse is terminated.

Figure 79 shows a circuit for a triode oscillator in which the pulse length is determined by an open-end delay network in the cathode circuit. Tube V_2 is biased beyond plate-current cutoff between pulses and is started by a pulse slightly wider than desired, applied to the driver V_1. The average oscillator cathode current initiates a pulse in the cathode delay network which is reflected from its open end with the same polarity and is returned to the cathode to raise its potential and terminate the pulse.

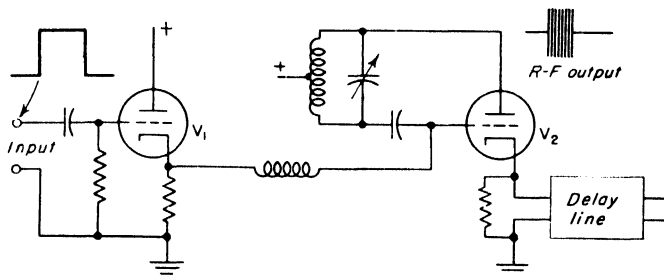



FIG. 79. Pulsed triode with delay line determining duration.

Alternatively, a delay network may be used in the grid circuit. In this case, a pulse of voltage negative relative to the grid is initiated by the grid current at the start of oscillation. The pulse travels to the end of the line and returns with the same polarity to bias the grid beyond plate-current cutoff and terminate the pulse.

THE CALCULATION OF DISTORTION

The calculation of amplitude distortion may be accomplished in at least three different ways. The first of these techniques, based on either the Fourier sum or, for nonrepetitive wave forms, the Fourier integral, is used with the output wave form itself in the time domain. The second is based on the Legendre polynomial expansion of the input-output relation of the device, and resolves the distortion in terms of orthogonal polynomials; it may be applied to devices having single-valued input-output relations, and also may be applied to devices having two outputs for each value of input as long as the load contour possesses central symmetry. (Such contours as ellipses and cyclic hysteresis loops have such symmetry.) The third is the use of some simple equations applying to many of the more common electronic circuits.

79. The Fourier Technique. The Fourier-series technique for the evaluation of harmonic content of a wave form takes advantage of the fact that trigonometric functions of different frequencies are orthogonal, that is, the integral

$$\int_0^{2\pi} \sin nt \sin (mt + \phi)\, dt \equiv 0 \tag{72}$$

holds identically when $n \neq m$ and when $\phi = 90$ deg. A corresponding integral involving cosines also holds exactly when $n \neq m$ and when $\phi = 90$ deg:

$$\int_0^{2\pi} \cos nt \cos (mt + \phi)\, dt \equiv 0 \tag{73}$$

As a consequence of these relations, any repetitive wave form may be resolved into harmonically related sine and cosine components which, when combined in an adding circuit, will reproduce the original wave form.

A resolution by Fourier analysis does not yield a unique solution because a simple phase shift of the wave form being analyzed will yield a completely new set of frequency components. For that reason, it is usually convenient to place the origin

of time at a point which will give either mirror symmetry or reversed mirror symmetry as in Fig. 80. If mirror symmetry can be used about the time origin, then only cosine functions need be considered; in reversed symmetry, only sine. If neither can be used, then both sine and cosine functions are required for the complete solution.

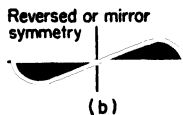
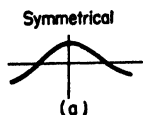



FIG. 80. Symmetry relations of periodic waves.

When analytical expressions for the waves are available, the integrals below may be used to find the Fourier coefficients a_n and b_n:

$$a_n = \frac{1}{\pi}\int_0^{2\pi} f(u)\sin nu\,du \tag{74}$$

$$b_n = \frac{1}{\pi}\int_0^{2\pi} f(u)\cos nu\,du \tag{75}$$

where the actual wave form when expanded in terms of the coefficients gives

$$f(u) = \frac{b_0}{2} + \sum_{j=1}^{\infty} b_j \cos jx + \sum_{j=1}^{\infty} a_j \sin jx \tag{76}$$

These expressions assume that the original wave has at most a finite number of discontinuities over the range $-\pi$ to $+\pi$ (or 0 to 2π); the value represented by the expression at the discontinuities is half the sum of the values on either side.

Fourier series do not always converge uniformly; for example, representations of square waves and rectangular-pulse waves provide only poor approximations to the function near the discontinuities. Convergence can sometimes be improved by the use of Fejer polynomials.[1]

FIG. 81. Square wave.

FIG. 82. Triangular wave.

FIG. 83. Saw-tooth wave.

80. Some Typical Expansions. Some of the more commonly useful Fourier expansions are listed below, together with their associated wave shapes:

Square wave (Fig. 81)

$$e = \frac{4E_0}{\pi}\left[\cos\theta - \frac{\cos 3\theta}{3} + \frac{\cos 5\theta}{5} + \cdots + \frac{(-1)^n\cos(2n+1)\theta}{2n+1} + \cdots\right] \tag{77}$$

Triangular wave (Fig. 82)

$$e = \frac{8E_0}{\pi^2}\left[\cos\theta + \frac{\cos 3\theta}{9} + \frac{\cos 5\theta}{25} + \cdots + \frac{\cos(2n+1)\theta}{(2n+1)^2} + \cdots\right] \tag{78}$$

Saw-tooth wave (Fig. 83)

$$e = \frac{2E_0}{\pi}\left[\sin\theta - \frac{\sin 2\theta}{2} + \frac{\sin 3\theta}{3} - \frac{\sin 4\theta}{4} + \cdots + (-1)^{n-1}\frac{\sin n\theta}{n} + \cdots\right] \tag{79}$$

[1] Guillemin, "The Mathematics of Circuit Analysis," p. 496, Wiley, 1949.

Fig. 84. Pulse train.

Fig. 85. Half sine wave.

Fig. 86. Rectified sine wave.

Clamp reference

Sine wave reference

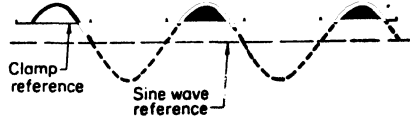



Fig. 87. Clamped sine wave.

Short rectangular pulse (Fig. 84)

$$e = E_0\left[k + \frac{2}{\pi}\left(\sin k\pi \cos\theta + \frac{\sin 2k\pi \cos 2\theta}{2} + \cdots + \frac{\sin nk\pi \cos n\theta}{n} + \cdots\right)\right] \tag{80}$$

Half-wave rectifier output (Fig. 85)

$$e = \frac{E_0}{\pi}\left[1 + \frac{\pi \sin\theta}{2} + \frac{2\sin 2\theta}{3} - \frac{2\cos 4\theta}{15} + \cdots + (-1)^{\frac{n}{2}+1}\frac{2\cos n\theta}{n^2 - 1} + \cdots\right] \tag{81}$$

Full-wave rectifier output (Fig. 86)

$$e = \frac{2E_0}{\pi}\left[1 + \frac{2\cos 2\theta}{3} - \frac{2\cos 4\theta}{15} + \cdots + (-1)^{\frac{n}{2}+1}\frac{2\cos n\theta}{n^2 - 1} + \cdots\right] \tag{82}$$

Biased diode (Fig. 87)

$$e = \frac{cE_0}{\pi}\left(\sum_{j=1}^{\infty}\left\{\frac{\sin[(n-1)\arccos a/c]}{n-1} + \frac{\sin[(n+1)\arccos a/c]}{n+1} - \frac{2a\sin[n\arccos a/c]}{nc}\right\}\right) \tag{83}$$

81. Graphical Analysis. The graphical analysis of wave forms is accomplished by dividing the basic period of repetition into $2m + 1$ equal increments of time and then establishing a series of algebraic equations, one for each of the selected time points, in terms of the harmonic components. For example, the value of the function or wave at the instant j may be written as

$$x_j = b_0 + \sum_{k=1}^{m} a_j \sin\frac{\pi jk}{m} + \sum_{k=1}^{m} b_j \cos\frac{\pi jk}{m} \tag{84}$$

where k is the order of the harmonic and j identifies the specific one of the $2m + 1$ increments of the period at which the data apply. Consequently, substitution of the appropriate values for x_j, $\sin(\pi jk/m)$, and $\cos(\pi jk/m)$ for the appropriate points in this equation and making one equation for each value of j gives a set of $2m + 1$ equations which can be solved for the coefficients a_k and b_k.

82. Legendre Polynomial Expansions. Distortion components can also be calculated from an amplifier input-output function or the input amplification function (the functions expressing the output or the amplification in terms of the input) by the use of Legendre polynomials, giving

$$f(s) = \sum_{j=0}^{\infty} c_j P_j(s) \tag{85}$$

The function $f(s)$ represents a set of values of the output voltage or the amplification at a series of discrete values s of the input voltage.

First the set of data being analyzed is evaluated at a series of equal increments of the input variable, the input voltage, and the standard technique is used for fitting a set of Legendre polynomials through the data representing the function. Briefly, the steps are:

1. Tabulate as a function of s the values of the input voltage e_i at which values of the function are available, where s is the number of increments displaced from the initial point.
2. Tabulate the function values $f(s)$ at the points s.
3. Write the values of the polynomials from the tables.[1]
4. Sum the squares of the polynomial values for given j as a function of the parameter s.
5. Sum the product of $f(s)P_j(s)$ for given j as a function of s.
6. Find the quotient of the values (5) divided by the corresponding values (4).

A typical analysis of the Legendre expansion of a function of amplification can be tabulated as in Table 5.

Table 5

e_i	s	$K(s)$	$P_0(s)$	$P_1(s)$	$P_2(s)$	$P_3(s)$	$P_4(s)$
e_{i0}	0	$K(0)$	$P_0(0)$	$P_1(0)$	$P_2(0)$	$P_3(0)$	$P_4(0)$
$e_{i0} + e$	1	$K(1)$	$P_0(1)$	$P_1(1)$	$P_2(1)$	$P_3(1)$	$P_4(1)$
$e_{i0} + 2e$	2	$K(2)$	$P_0(2)$	$P_1(2)$	$P_2(2)$	$P_3(2)$	$P_4(2)$
$e_{i0} + he$	h	$K(h)$	$P_0(h)$	$P_1(h)$	$P_2(h)$	$P_3(h)$	$P_4(h)$
S_j			$\sum_{h=0}^{n} P_0^2(h)$	$\sum_{h=0}^{n} P_1^2(h)$	$\sum_{h=0}^{n} P_2^2(h)$	$\sum_{h=0}^{n} P_3^2(h)$	$\sum_{h=0}^{n} P_4^2(h)$
B_j			$\sum_{h=0}^{n} K(h)P_0(h)$	$\sum_{h=0}^{n} K(h)P_1(h)$	$\sum_{h=0}^{n} K(h)P_2(h)$	etc.	
C_j			$P_0(n)B_0/S_0$	$P_1(n)B_1/S_1$	$P_2(n)B_2/S_2$	etc.	

As can be seen, the C_js are the coefficients of the polynomials in Eq. (85). The values of the P_j functions are tabulated in many books on numerical analysis in addition to Milne; the values for a few cases are given in Table 6. The data have been reduced to corresponding integral values, since the divisors can be shown to cancel out in the process of analysis.

83. Calculations with Legendre Polynomials. The Legendre coefficients may be converted directly into the amplitudes of the harmonic components by the use of conversion equations.

The set for converting the data on amplification as a function of input bias is

$$\begin{aligned} E_5 &= K_4/80 = 35C_4/640 \qquad E_4 = K_3/32 = 5C_3/64 \\ E_3 &= -K_2/12 - K_4/16 = (15C_4 - 48C_2)/384 \\ E_2 &= -(2K_1 + K_3)/8 = (C_3 - 4C_1)/16 \\ E_1 &= K_0 + K_2/4 + K_4/8 = C_0 - C_2/8 - C_4/64 \end{aligned} \tag{86}$$

where the various E_j apply to the equation

$$e = E_0 + E_1 \sin \omega t + E_2 \sin 2\omega t + E_3 \sin 3\omega t + \cdots + E_j \sin j\omega t + \cdots \tag{87}$$

and the C_j are obtained from the equation

$$K = \sum_{j=0}^{m} C_j P_j(s) \tag{88}$$

[1] See the included tables and also Milne, "Numerical Calculus," p. 375, Princeton University Press, 1949.

Table 6. Table of Values of Legendre Polynomials

$n = 4$						$n = 6$					
P_0	P_1	P_2	P_3	P_4	s	P_0	P_1	P_2	P_3	P_4	P_5
+1	−2	+2	−1	+1	0	+1	−3	+5	−1	+3	−1
+1	−1	−1	+2	−4	1	+1	−2	0	+1	−7	+4
+1	0	−2	0	+6	2	+1	−1	−3	+1	+1	−5
+1	+1	−1	−2	−4	3	+1	0	−4	0	+6	0
+1	+2	+2	+1	+1	4	+1	+1	−3	−1	+1	+5
					5	+1	+2	0	−1	−7	−4
					6	+1	+3	+5	+1	+3	+1
5	10	14	10	70	S_j	7	28	84	6	154	84

$n = 8$						$n = 10$				
P_1	P_2	P_3	P_4	P_5	s	P_1	P_2	P_3	P_4	P_5
−4	+28	−14	+14	−4	0	−5	+15	−30	+6	−3
−3	+7	+7	−21	+11	1	−4	+6	+6	−6	+6
−2	−8	+13	−11	−4	2	−3	−1	+22	−6	+1
−1	−17	+9	+9	−9	3	−2	−6	+23	−1	−4
0	−20	0	+18	0	4	−1	−9	+14	+4	−4
+1	−17	−9	+9	+9	5	0	−10	0	+6	0
+2	−8	−13	−11	+4	6	+1	−9	−14	+4	+4
+3	+7	−7	−21	−11	7	+2	−6	−23	−1	+4
+4	+28	+14	+14	+4	8	+3	−1	−22	−6	−1
					9	+4	+6	−6	−6	−6
					10	+5	+15	+30	+6	+3
60	2,772	990	2,002	468	S_j	110	858	4,290	286	156

The columns P_0 have been omitted for $n = 8$ and $n = 10$ since each value in the column is positive unity, and the sum S_j is $n + 1$.

The set of equations for converting the data on output voltage as a function of the input voltage is:

$$\begin{aligned} E_5 &= 63C_5'/128 \qquad E_4 = 35C_4'/64 \qquad E_3 = -5C_3'/8 - 35C_5'/128 \\ E_2 &= -3C_2'/4 - 5C_4'/16 \qquad E_1 = C_1' + 3C_3'/8 + 15C_5'/64 \\ E_0 &= C_0' + C_2'/4 + 9C_4'/64 \end{aligned} \tag{89}$$

where the E_j are defined in Eq. (87), and the C_j' are the coefficients of the equation

$$e_0 = \sum_{j=0}^{n} C_j' P_j(s) \tag{90}$$

These two sets of equations give the amplitudes of the respective harmonic components in terms of the coefficients of the corresponding Legendre polynomials.

84. Bivalued Contours. A similar technique may be applied to elliptical or other load lines having central or radial symmetry, giving both the in-phase and the quadrature components of the various harmonics present in the output wave. The analysis is divided into two parts, (1) the calculation of the equivalent in-phase components, corresponding roughly to the magnetization curve with the hysteresis loop (Fig. 88), and (2) the reduction of the asymmetrical excursions into quadrature components. With Legendre polynomials, the first operation is to fit a fifth-order set of polynomials to the contour or load line over one-half of its cycle, such as branch

A in Fig. 88. The symmetrical portion of this set of polynomials consists of the odd-order group (based on the input-output relation as the contour). The even-order polynomials may then be replotted, using one extreme point (e.g., *C*) as the center, and taking a reverse polarity or sign for the polynomial values on one side of the new center point (Fig. 89). The resulting contour may then be reanalyzed in terms of odd-order polynomials. In this analysis, however, a new set of data points must be taken, these points being related to the set used in the first step in accordance with Table 7, where the $\Delta e = (E_{cp} - E_{cs})/n$, with E_{cp} being the most positive bias voltage in the cycle, E_{cs} the static value of voltage, and n being the total number of data points used per quadrant. Use of the listed data points assures uniform spacing and proper symmetry for the replotted quadratic terms.

Fig. 88. Hysteresis loop.

Fig. 89. Components of bivalued input-output relation.

Table 7. Table of Data Points for Reduction of Elliptical Contours

Symmetric points		Asymmetric points	
Angle	Voltage	Angle	Voltage
arcsin 0	0	arccos 0	$n\Delta e$
arcsin $1/n$	Δe	arccos $1/n$	$\Delta e \sqrt{n^2 - 1}$
arcsin $2/n$	$2\Delta e$	arccos $2/n$	$\Delta e \sqrt{n^2 - 4}$
arcsin $3/n$	$3\Delta e$	arccos $3/n$	$\Delta e \sqrt{n^2 - 9}$
.	.	.	.
.	.	.	.
.	.	.	.
arcsin 1	$n\Delta e$	arccos 1	0

The correctness of the results may be readily verified by checking through the expansion of any bivalued function having central symmetry. The functions to be summed in terms of phase angles are:

$$E_1 \sin t$$
$$E_1' \sin (t + 90°)$$
$$E_3 \sin 3t$$
$$E_3' \sin (3t + 90°)$$
$$E_5 \sin 5t$$
$$E_5' \sin (5t + 90°)$$

where the unprimed *E*s are the amplitudes of the trigonometric components from the symmetric or odd polynomials, and the primed *E*s are the amplitudes of the trigonometric components from the asymmetric or even polynomial components after they have been converted to odd-order polynomials.

85. Equations for Distortion. In addition to the analytic techniques for evaluating distortion, techniques for rapid approximation of amplifier distortion have been used for many years. Some of these equations are expressed in terms of averaged amplification (strictly a voltage change), and others in terms of the actual small-signal amplifications. Where the required data are available, the expressions in terms of

small-signal amplification give the user more complete understanding of the problem, with the result that they are somewhat more satisfactory.

Second Harmonic Distortion. The approximate equation in terms of averaged amplification or voltage change is

$$D = 50(E_{bn} + E_{bp} - 2E_{bs})/(E_{bn} - E_{bp}) \tag{91}$$

In this equation, E_{bn} is the plate voltage at which the grid voltage is most negative, E_{bp} the plate voltage at which the grid is most positive, and E_{bs} the static plate voltage. The distortion D given in Eq. (91) is the approximate percentage of second harmonic generated.

In terms of small-signal amplification, Eq. (91) may be written in the form

$$D = 25(K_p - K_n)/(K_p + K_n) \tag{92}$$

In this equation, K_p is the amplification at the positive bias limit, and K_n the amplification at the negative bias limit.

Third Harmonic. If the distortion is primarily third harmonic, as it is, for example, in a properly balanced push-pull amplifier, the determination of the distortion requires data on the amplification at the static bias point (K_s) in addition to the two limit points. The equation for the third harmonic distortion with $K_p = K_n$ then is

$$D = 100(K_s - K_p)/3(3K_s + K_p) \tag{93}$$

If K_p and K_n are nearly, but not exactly, equal, then the equation reads

$$D = 100(2K_s - K_p - K_n)/3(6K_s + K_p + K_n) \tag{94}$$

MISCELLANEOUS NONLINEAR CIRCUITS

86. Nonlinear Elements in AVC and AGC. Amplifiers which amplify a wide range of signal voltages and provide a relatively constant output signal are expected to do so with a minimum of distortion. A smooth variation of amplification as a function of a control voltage is the characteristic required of the tube used in this application. The ratio of the change in amplification to the average amplification should be proportional to the signal voltage and relatively independent of the static bias applied. Or $\Delta K/K_0 = c\Delta e_c$, and integrating, $\log (K/K_0) = ce_c$, where K is the amplification at the bias e_c, K_0 is the maximum amplification, and c is an arbitrary constant.

Although amplifiers having logarithmic characteristics are by no means distortionless, on the average they can provide the required variation in amplification with comparatively small distortion. Tubes which are used for logarithmic amplifiers are called *remote-cutoff* tubes. Although pentodes such as the 6BA6, the 6BJ6, and the 6SK7 usually are used in this service, some dual triodes for use as cascode amplifiers have recently been introduced. The 6386 is such a tube.

The variable-gain amplifier is normally used in r-f or i-f circuits of communication equipment. The different values of amplification are selected by the use of an AVC circuit, which meters the incoming signal voltage and returns a bias to the amplifiers to hold the average output voltage reasonably constant. To prevent excessive distortion, usually only a very small part of the AVC control signal is applied to the final amplifier. As a result, most of the control is obtained in the early stages of the receiver.

87. AVC Control Circuits and Amplifiers. The circuit providing the control for an AVC amplifier may return a voltage which is approximately proportional to the average signal at the detector, or it may have no effect until a certain chosen signal level is reached (*delayed AVC*). Best results with delayed AVC are obtained if a separate, constant-gain, final amplifier is used to provide the control signal to the AVC detector; with a proportional AVC system, the signal detector may be used for AVC as well as signal detection if desired. The separate amplifier is required to prevent clipping distortion from being generated on the modulation.

Additional amplification, either before or after detection, may be used with delayed AVC to hold the output-signal variation to a minimum. The isolating amplifier can

also contribute substantially to improving the constancy of gain. Since a voltage gap of 1 or 2 volts minimum is desirable to provide for effective delay on the AVC action, the extra amplification is usually included in the i-f portion of the system. A d-c amplifier can be helpful after detection if a suitable negative supply is available to power it. Average output-signal-level variations of as small as 10 per cent may result with a change of input-signal level of 40 to 60 db in properly designed AVC amplifier systems.

88. Design of Amplifiers. The design of variable-gain amplifiers, because of the amplification range over which the coordination of static and small-signal data is required, is best accomplished with the help of a logarithmic type of screen-characteristic data sheet for the tube which includes small-signal data. A sheet such as

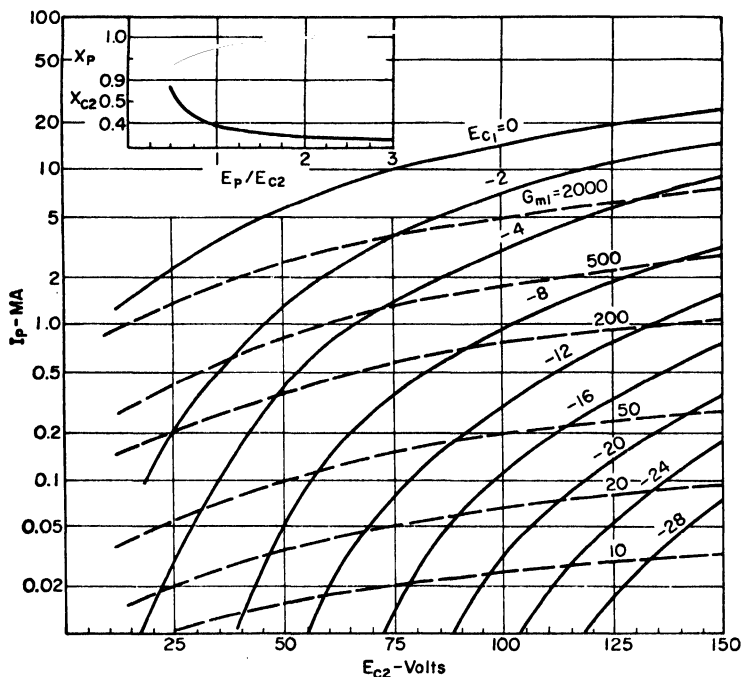



FIG. 90. Screen characteristics, 6BJ6.

that shown in Fig. 90 has proved useful for the application, since the range of bias and small-signal characteristics required to design the amplifier is available in addition to an adequate distribution of the data on the plate current. The transconductance contours are plotted on a decade basis to simplify the calculations at the widely varying conditions encountered.

In the design of an i-f amplifier, the steps required after the selection of a suitable tube with its appropriate characteristic data include (1) the selection of the screen voltage to be used (or the minimum screen voltage if a series screen supply such as that shown in Fig. 91 is used), (2) the selection of the maximum amplification desired per stage (usually 20 or less where stability is of prime importance), and (3) the design of the required AVC supply circuits. A typical circuit is shown in Fig. 91. After the screen voltage has been selected, a screen load line may be plotted; it is represented by a vertical line at the appropriate voltage if a fixed screen supply is used, or it may be calculated from the equation

$$e_{c2} = E_{cc2} - I_p X_{c2} R_s \qquad (95)$$

where E_{cc2} = screen-supply voltage
e_{c2} = screen voltage
I_p = nominal plate current
X_{c2} = screen-current-division factor (see Sec. 15)
R_s = series resistance between screen and screen supply

The values of I_p and X_{c2}, in addition to X_p and G_{m1}, may be read directly from the screen-characteristic curve sheet. After these data are tabulated, the characteristics of the amplifier stages as a function of the control voltage may be calculated, after which the over-all characteristics of the amplifier may be obtained by selecting a series of output voltages, finding the corresponding control voltage, and calculating successively the inputs to the individual amplifier stages, starting at the output. The distortion may be checked at the same time. The gain of the individual amplifier stages may be found from the equation

$$K = -G_{m1}X_pZ_L \qquad (96)$$

where Z_L is the tuned impedance and G_{m1} and X_p are obtained from the curve sheet.

89. Linear Time Variant (LTV) Circuits. Circuits in which the variations in small-signal parameters are a function of some input or control signal other than the principal signal being processed are not in the strictly mathematical sense nonlinear

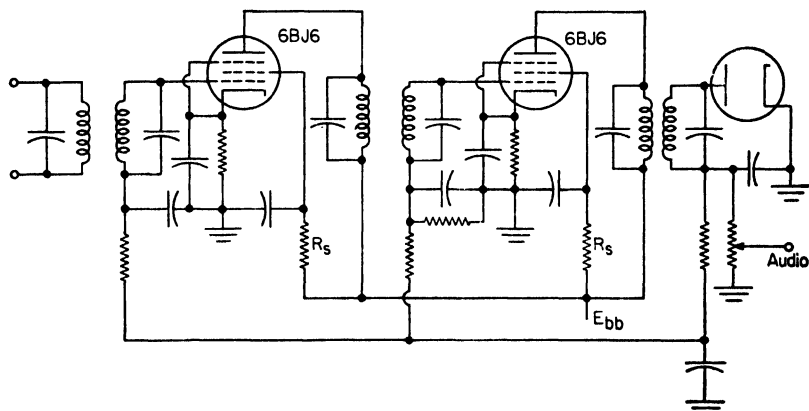



Fig. 91. Typical i-f amplifier.

although they behave in a manner similar to circuits which are nonlinear. The behavior is described as linear time variant (LTV) since, if the control signal, which is usually time-dependent, is held at a constant value, the operation of the amplifier would be completely linear. On an instant-by-instant basis, therefore, the theorem of superposition and other criteria governing the action of linear circuits are applicable to LTV circuits.

The variation of the circuit characteristics is usually obtained through the use of a control voltage or current which is periodic in nature (often a sine wave). This control signal must have an amplitude which is large compared with that of a signal whose characteristics it is altering in step with the control. The result is that an LTV circuit acts as an electronic multiplier, multiplying the received signal by the control signal alone without permitting components of the received signal to be multiplied by each other.

90. Mixers as LTV Circuits. Mixer circuits are probably the most commonly used form of circuit having LTV characteristics. They possess the LTV property as long as the received signal is small in amplitude in comparison with the local oscillator signal. If the received signal can be amplified by the mixer with negligible distortion in the absence of the local oscillator signal, adequate linearity is available

to assure LTV operation. Although operating conditions usually are such that the assumption is valid, in the presence of either noise of excessively high amplitude or a strong interfering signal, the behavior may not be linear time variant with a mixer.

91. The Autocorrelation Detector. The autocorrelation detector is a form of LTV circuit using a mixer as a multiplier in which the received signal is divided into two parts and then introduced after separate processing to the two multiplier inputs. One of the parts of the signal goes directly to one input, and the other is passed through a delay circuit which delays the signal and noise for a sufficiently long period of time to permit the noise in the two parts of the signal to become random with respect to one another prior to its introduction to the second input. The components of the desired signal, because of its periodic nature, still maintain a coherent relation in spite of the delay, with the result that the noise tends to cancel, whereas, with correctly selected delay, the components of the desired signal reinforce one another. A typical circuit is shown in Fig. 92.

92. The Cross-correlation Detector. The cross-correlation detector differs from the autocorrelation detector primarily in the form of the signals used in the input channels. Two signals, the received signal and a processing signal, are applied to the two inputs of the LTV circuit. The characteristics of the received signal are usually known only in general form rather than in detail. The second, or locally generated signal, either may be very similar in nature to the desired signal, in fact a

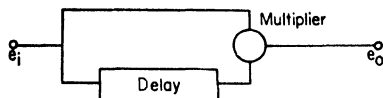



FIG. 92. Autocorrelation circuit.

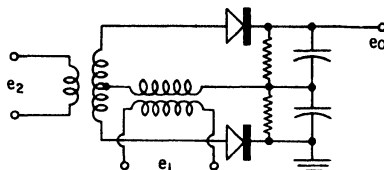



FIG. 93. Phase detector.

complementary signal if the equipment designer can produce one, or it may have characteristics which will facilitate recognition of the received signal. Often all that is required of the local signal is that its frequency be as nearly as possible exactly the same as that of the carrier of the received signal; if, however, the intelligence is actually conveyed by the carrier itself instead of the modulation, then a frequency offset of a few hundred or a few thousand cycles may provide better operating conditions.

One of the commonly used forms of cross-correlation detector is the phase detector (Fig. 93). This device can be used to keep two frequencies exactly in step, and in fact to maintain quadrature between the received signal and the local signal to within a few electrical degrees. The d-c output from the phase detector may be used both as a control voltage to keep the correct phase relation and also to give a signal which is proportional to the frequency deviation (or modulation). Further discussion of the action may be found in Sec. 60.

93. Regenerative Cross-correlation Detector. The regenerative detector, when used with weak input signals, behaves similarly to a cross-correlation detector because of the control exerted by the local oscillations. As the circuit is ordinarily built (Fig. 31) the variable grid loading makes operation somewhat nonlinear, with the result that the inherent characteristics of the detector often are not adequately utilized. If, however, circuits such as those shown in Fig. 32*a* and *b* are used, then the operating characteristics with small signals are rather linear, with the result that the high Q of the tuned circuit may be utilized effectively, and the circuit behaves largely as a cross-correlation detector. The local oscillation actually serves two purposes, first to build up the effective Q of the tuned circuit to produce, in effect, linear passive amplification of the signal, and to introduce correlation detection after the amplitude is sufficiently large.

94. Signal-to-noise (S/N) Characteristics of Detectors. One of the properties of particular concern in any type of detector, whether it is piecewise-linear, nonlinear, or time-variant, is its ability to separate the desired incoming signal from the combination of signal and noise. Representing symbolically the combination of signal and noise as $[S] + [N]$, the output from either the piecewise-linear detector or the nonlinear detector takes the form.

$$[S][S] + [S][N] + [N][S] + [N][N] \tag{97}$$

and can be reduced to the form

$$[S][S] + 2[S][N] + [N][N] \tag{98}$$

If this equation is reorganized to give S/N, it takes the form

$$S/N = [S]/(2[N] + [N][N]/[S]) \tag{99}$$

since both the last two terms in Eq. (98) constitute noise terms. As can be seen, the noise components in the denominator have at least twice the amplitude they have in the incoming signal. If the signal is appreciably stronger than the noise, then the signal-to-noise ratio is cut in half, whereas if the noise is large compared with the signal, then the signal will be embedded deeply in the noise. Consequently, nonlinear types of detectors are poor weak-signal detectors.

95. S/N Characteristics of Autocorrelation Detector. With the autocorrelation detector, an action similar to ordinary detection takes place, the one difference being that the noise carried with the two signals being multiplied in the detector differs appreciably at the two inputs because of the time delay introduced in one channel. If the delay is sufficient, the noise in the two channels becomes uncorrelated, with the result that Eq. (98) becomes

$$[S]([S] + [N] + [N^0]) + [N][N^0] \tag{100}$$

Now, since $[N]$ and $[N^0]$ are uncorrelated, their sum is not $2[N]$ but a number less than it. On the average, statistically, the sum with random noise has the value $\sqrt{2}\,[N]$; at the same time the term $[N][N^0]$ will have a value less than $[N][N]$ because of the statistical averaging. Approximately 3 db of improvement in signal-to-noise ratio is obtained by the use of an autocorrelation detector in place of a nonlinear detector.

96. S/N Characteristics of Cross-correlation Detector. The cross-correlation detector requires a somewhat different form of equation than the autocorrelation because of the locally generated control signal. Identifying the control signal as $[S^0]$, the output takes the form

$$[S][S^0] + [N][S^0] \tag{101}$$

Unless noise or modulation is present on $[S^0]$, these are the only two terms present. Consequently, the output signal-to-noise ratio is

$$(S/N) = [S]/[N] \tag{102}$$

giving a post-detection signal-to-noise ratio which is identical with the predetection ratio; the cross-correlation detector, when used with a noise-free control signal $[S^0]$, then provides an output which is optimum as a signal-separation circuit.

The post-detection S/N for the cross-correlation detector, for band widths wider than the spectral width of the received signal, is inversely proportional to the square root of the band width, just as is the predetection ratio. On the other hand, nonlinear or piecewise-linear detectors have a post-detection S/N ratio which is inversely proportional to the fourth root of the band width, or definitely poorer than the predetection ratio. For this reason, the use of cross-correlation detection is indicated whenever quality is of greater importance than price in electronic equipment.

The significance of the data in the above paragraph is verified by the following results, which have been obtained in the microwave spectrum. King has reported

that well-designed superheterodyne receivers operating at K band have a sensitivity of approximately −130 dbw (db below 1 mw) with a conventional detector, −135 dbw with a rotating wave-guide phase shifter, and −140 dbw with a correlation, or synchronous, detector. The 10-db difference between the straight detector and the correlation detector as reported is consistent with the estimate of a minimum of 6-db improvement using correlation techniques. The phase shifter improves the characteristics somewhat by making the system partially time-variant and evidently makes the signal somewhat easier to recognize in the presence of noise.

Further data from the same source show that the use of a nonlinear detector directly without heterodyning and amplification is even poorer. If a 1N26 diode detector is used directly as a detector with a 4-cps audio amplifier, the system sensitivity is −100 dbw, and with a bolometer under similar conditions −90 dbw. These data show the low conversion or detection efficiency of the two detectors and their relatively high noise levels.

97. Noise-integration Techniques. Advantage can be taken of the lack of correlation in noise as compared with the desired signal to provide a random summation of the noise, and linear summation of the signal, thereby improving the over-all

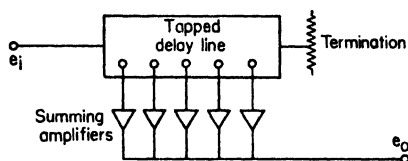



FIG. 94. Transversal filter.

signal-to-noise ratio beyond the theoretical ideal obtainable without integration. Any type of time-delay filter may be used for noise integration by tapping the filter at equally spaced time delays, these delays corresponding to either half the period of the signal frequency or the full period. The use of a series of summing amplifiers at the respective taps, properly combined, causes the components of the desired signal to add at the output and the noise to tend to cancel (Fig. 94).

The reduction of noise may be understood by reference to the statistical problem of random walks, which states that, if n different random elements are added together, the probable average amplitude of the combination at the output is not the sum of the elements, but approximately

$$\bar{A} = \sqrt{n}\,\bar{x} \tag{103}$$

where n is the number of summations and $\bar{x}$ is the average magnitude of the random elements without regard to polarity (the average of the absolute values).

The time delay, or transversal filter,[1] as it is sometimes called, is only one of a number of devices which can be used for noise integration. Among some of the other techniques are cathode-ray integration, magnetic-tape or drum integration, etc. All these function by adding the coherent signal in phase and the noise in its own random phase, giving as a result an improved signal-to-noise ratio. Such devices are sometimes called *comb filters*.

[1] Kallman, H., The Transversal Filter, *Proc. IRE*, July, 1941.

CHAPTER 13

DIRECT-CURRENT AND LOW-FREQUENCY MEASUREMENTS

BY JOHN H. MILLER[1]

STANDARDS

1. General. True basic measurements of electrical quantities are rarely made except in standardizing laboratories, owing to the inherent difficulties in the procedure. Ordinary measurements are made by comparison devices of one form or another. Direct-reading instruments, having an electrical torque-producing means functioning against a spring, are calibrated against accurate standards which are in turn calibrated against basic measuring devices. Such torque-producing instruments are used for measuring current, voltage, power, and resistance. Instruments for measuring phase relations, frequencies, and other factors may have two torque-producing systems, each torque varying with the position of the moving element and bearing different functional relations to the quantity measured. The result is for the moving system carrying the pointer to take up a position where the torques balance, this being different for each different value of the quantity in question, and the scale may be marked accordingly.

Electrical units such as the volt, the ampere, and the ohm are based upon and are intended to be exact multiples of the units of the cgs electromagnetic system. The international units which were made standard in 1911 were derived according to the best procedures then available. However, refinement of measurement techniques in the years following, along with intercomparison of the units by the national laboratories of various countries, indicated that the international volt and ohm were larger than their absolute values by 0.0495 and 0.033 per cent, respectively. Accordingly, on Jan. 1, 1948, the absolute values were made standard and form the basis for all electrical measurements and instrument calibrations since that date.

2. Current. Current is measured, absolutely, in terms of the force of attraction or repulsion between two coils connected in series and carrying that current. More specifically, one cgs unit of current (1 abampere or 10 amp) flowing in a circuit consisting of two infinitely long parallel conductors causes a force of 2 dynes per cm length to act between them when they are 1 cm apart. From this basic definition to the particular case of the force between two coils in a current balance requires an integration of the total force of each element of each coil on the opposite coil. In turn this means extremely refined measurement of the several dimensions involved, *i.e.*, effective coil diameter, wire diameter, coil spacing, etc. The force developed between the coils for a given current is determined by balancing with known weights and in terms of the existing gravitational constant.

The use of the standard current balance is tedious and time-consuming, and determinations with it are made in the national laboratories only occasionally.

Actually the current balance as described is used for the calibration of a standard cell by balancing the voltage drop of the measured current through a known resistance against the potential of the cell. The standard of current is thus maintained through the application of Ohm's law, a standard of resistance, and the voltage of a standard cell.

[1] Weston Electrical Instrument Corp., a subsidiary of Daystrom, Inc.

3. Resistance. The ohm was defined in 1893 as the resistance of a mercury column 1 mm square and 106.3 cm long at 0°C. The ohm is presently determined in absolute terms by balancing the voltage drop of a current through the resistance in question against the rectified result of the a-c voltage induced in one coil of a mutual inductance, where the same current is commutated into the other coil. Balancing is accomplished by varying the commutating frequency. Thus the resistance is evaluated in terms of mutual inductance—a matter of coil dimensions—and frequency—a function of time. The standard ohm is now maintained by the National Bureau of Standards as the average of a group of 1-ohm manganin resistors, these units having been checked for their absolute values as described and having shown a change of less than 1 ppm from the mean of the group over many years. Practical secondary standards are coils of manganin wire immersed in oil and sealed in metal containers. Such sealed standards built by Leeds & Northrup to the specifications of the National Bureau of Standards are adjusted to an accuracy of 0.01 per cent and may be relied upon to hold their calibration to 1 part in 100,000 for considerable periods of time. The sealing of the containers is important to prevent the absorption, by the oil, of moisture from the atmosphere, for such moisture will deposit upon the shellac or other insulating material on the wire which, in turn, will cause mechanical strains to distort the values beyond normal expectancy.

4. Voltage. Voltage measurements cannot be made absolutely with an accuracy sufficient to make the measurement desirable, on account of the smallness of the electrostatic forces involved. The standard of voltage is maintained by the National Bureau of Standards as the average of a number of Weston *saturated* cadmium cells, sometimes known as the "Weston normal cell." The average value of this bank of cells is 1.018636 absolute volts at 20°C, and this value is cross-checked from time to time through the use of the standard ohm and the current balance as described above. These cells, as built by Weston and Epley, are correct to 0.001 per cent of the values given. The cell has a small but appreciable temperature coefficient, and in use its temperature must be maintained constant at the specified value. For this reason the saturated cell is not well adapted for commercial laboratory use.

The *unsaturated* cadmium cell, the type of standard cell customarily used for laboratory work, has a negligible temperature coefficient but must be compared with the saturated type for its initial calibration. Its voltage is constant to better than 1 part in 10,000, but the cells must be recertified yearly as the potential tends to drop very approximately 50 μv per year. Standard cells must always be used in a null fashion with no more than 50 μa drawn from the cell at any time.

5. Reactance. The self and mutual inductance of single-layer air-core coils and the capacitance of two-plate air capacitors having guard rings may be calculated from their dimensions, with an accuracy of better than 2 parts in 100,000. The standard of capacitance is maintained by the National Bureau of Standards as a bank of quartz-insulated air capacitors.

6. Frequency. The absolute standard of frequency is the mean solar day as measured by astronomical observations. At the present writing the International Committee on Weights and Measures is expected to adopt a definition of the second, as a unit of time, as "the fraction 1/31,556,925.975 of the tropical year 1900." The mechanical vibrations of quartz crystals provide standards of frequency when permanently connected into suitable vacuum-tube circuits and allowed to oscillate continuously at constant temperature. Their frequency is constant to 1 part in 10^9 per month and 1 part in 10^{10} per day for the best quartz standards. A variety of shapes of crystals have been used to attain freedom from the effect of minor temperature changes as well as changes in the constants of the associated electronic gear. Even with the best control of temperature and voltage there are short-time fluctuations of the order of 1 part in 10^{11}. The frequency of the crystal with which such accuracy may be obtained is near 100 kc for low-frequency types and 1 to 10 Mc for high-frequency types. By means of suitable multipliers and dividers all other frequencies may be obtained with the same accuracy.

Tuning forks made from carefully stabilized metals also provide frequency standards in conjunction with suitable tube circuits but usually of a lower accuracy; the best stability which has been reported is 1 part in 10^6 for several weeks.

Mention should also be made of the use of the constant properties of atoms and molecules as standards of frequency. For example, the use of the 3,3 inversion line of the ammonia molecule in the feedback loop of a stable oscillator will control the frequency to a value given by Shimoda as 23,870,131 kc. A cesium atomic beam has also been used in similar fashion.

Quartz crystals whose frequencies remain constant to better than 1 part in 10^6 may be made in the frequency range 20 kc to 75 Mc. Metals, such as nickel and certain iron alloys, having the property of magnetostriction, may be used as oscillators in suitable tube circuits. Their frequency range extends from 5 to 100 kc. Their stability is about 2 parts in 100,000. For the lower frequencies tuning forks and metal bars are used. Their frequency range is 25 to 1,000 cycles.

Reference is made to the summary paper by F. D. Lewis, Frequency and Time Standards, *Proc. IRE*, September, 1955, p. 1046. This paper includes a bibliography of 74 pertinent references.

CURRENT-MEASURING INSTRUMENTS

In the following discussion of electrical measuring instruments as such, and particularly of panel, switchboard, and portable types, a valuable reference is the American Standard, C39.1-1955, entitled "Electrical Indicating Instruments," obtainable from the American Standards Association, 70 East 45th Street, New York 17, New York. It is a very complete document with definitions of terms and standard requirements for the various classes of instruments which are adhered to by most makers.

Current-measuring instruments are discussed here in detail since most instruments basically respond to the level of current flow in their windings. Voltmeters are simply sensitive milliammeters with resistance units in series with the terminals and of such value that, at the marked full-scale potential, full-scale current flows through the resistance and actuating windings.

7. Moving-coil permanent-magnet instruments of the pointer type, or reflecting galvanometers, consist of a coil, usually wound on a metal frame for damping purposes, which can rotate in an intense uniform magnetic field produced by a permanent magnet.

The current I flowing through the turns N of the coil reacts with the magnetic field H in the air gap to produce a force F acting on each conductor proportional to the product IHl of the current, magnetic field, and length of conductor in the field. If the coil is pivoted at its center, a torque will be exerted, tending to rotate the coil about an axis parallel to the sides of the coil and perpendicular to the magnetic field. Some kind of restoring torque is provided which is proportional to the angle θ through which the coil rotates. Expressing the sensitivity S of the instrument as the angular deflection per unit current, it is given by

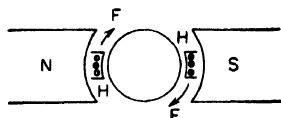



FIG. 1. Moving-coil galvanometer.

$$S = \frac{\theta}{I} = \frac{HNlb}{\tau} \tag{1}$$

where b is the diameter of the coil and τ is the restoring torque per unit angular displacement.

For maximum sensitivity a high flux density is desirable, although the effective upper limit appears to be about 4,000 gausses; higher values may act on otherwise imperceptible magnetic inclusions in the moving system and cause additional random torques which will give nonlinear deflectional characteristics. Further, since damping is a function of the square of the flux, high flux values may give overdamped deflectional characteristics even without a damping frame to the point where circuit damping

alone may require many seconds for a final reading. Sensitivity increase by flux increase alone is, therefore, of limited usefulness.

8. Torque to Weight Ratio. The torque should be as low as possible for high sensitivity. Suspended-type instruments show no friction, but low torque lengthens the period. Again there is a practical lower limit. In pointer-type instruments where the moving element rotates on pivots between sapphire or the more recent high-silica glass V jewels, there is a minimum torque which may be used for a given moving element weight in order that frictional effects will be unobservable. For instruments mounted on a switchboard and having a horizontal axis, the ratio of the full-scale torque in milligram-centimeters with the weight in grams should not be less than 40 for small instruments, 60 for larger instruments of 1 per cent accuracy, and still greater if greater accuracy is required. For portable instruments having a vertical axis, it has been found that heavy elements, over 1 g, show greater friction than given by the above relation, and lighter elements show less friction. Hence for such vertical axis instruments for portable service the torque/weight$^{3/2}$ ratio is used and this ratio should be over 40 for small instruments and over 60 to 100 for large instruments for unobservable friction. Ratios much lower than this may be satisfactory for highly sensitive laboratory instruments used with care and not subject to vibration or handling.

9. Permanent Magnets. The magnetic field obtained from the permanent magnet must be constant so that the electrical characteristics of the instrument may remain unchanged. The constancy of a magnetic system is determined by the ratio K, which is equal to the product of the effective length of the magnet times the effective cross section of one of the air gaps, divided by the product of the cross section of the magnet and the total air-gap length. This constant should be over 100 for chrome and tungsten magnet steels and over 30 for high-cobalt steels. For the various nickel-aluminum or MK steels the constant will vary, but 10 may be taken as a median value.

Tungsten and chrome steels are still used to some extent, and although they result in relatively bulky structures, they still represent an inexpensive type of magnet for moderate flux densities. High-cobalt steels have been used but they are costly and have generally been replaced by various types of Alnico. Alnico V, in particular, both gives high flux density and can be used in a short length; its total energy content—the product of internal flux density and coercive force—is presently the highest available.

FIG. 2. Magnetic system (a) and cutaway view (b) of core-magnet mechanism.

Since the coercive values are very high, short magnetic systems are possible and in recent years the use of a core-magnet system has become quite prevalent. Figure 2a shows one form of such an arrangement where the internal magnet has sufficient energy to drive an adequate amount of flux through the gap to the surrounding ring structure and results in a very compact instrument mechanism as in Fig. 2b. Available structures give gap densities of 1,500 to 2,500 gausses in the air gap in conventional panel, switchboard, and portable instruments; in special structures where the magnet length can be somewhat increased, higher densities may be obtained.

An important quality of the core-magnet arrangement is its high degree of flux utilization. All the leakage flux is useful because the magnet is completely surrounded

by the moving coil, whereas the leakage flux with an outside magnet may be as much as 50 per cent. This again results in a smaller and more compact structure of lighter weight. Another important quality is that of self-shielding since the outer soft-iron ring or yoke not only serves as a return path for the main flux but also diverts extraneous flux around the air gap. More specifically, in a stray field of 20 oersteds which might be obtained from a conductor 1 ft away carrying 3,000 amp or, indeed, from an adjacent magnetron, an older model 1 portable instrument might be influenced as much as 10 per cent of its reading whereas with the new core-magnet structure the effect would be less than 0.3 per cent.

Where very high flux densities are needed they can be obtained with a massive external magnet. In special cases, densities as high as 9,000 gausses have been obtained over a narrow angle.

Where a distorted d-c scale is required to balance other factors, for example, to give an evenly divided scale marked in decibels, it is possible to cut the pole tips to produce a markedly distorted field resulting in a more useful scale for the quantity measured. It is more difficult, however, to manufacture such nonuniform pole pieces so that such instruments are inherently more expensive and the basic accuracy is usually somewhat less; cut-pole-piece instruments are quite widely used, however, because of the useful end result on the scale, and when hand-drawn scales are furnished the accuracy is usually adequate.

10. Deflection Indicators. The deflection of any sensitive galvanometer is indicated by the angular rotation of a beam of light, the so-called *optical lever*, which is reflected from a mirror, either plane or convex, mounted above the moving coil. The older form of telescope and scale is now being replaced by a spot of light containing cross hairs which moves along a scale. The use of a spot of light is much less fatiguing than observation through a telescope, and a wider range of view is obtained. The usual scale length is 50 cm with zero in the center. The standard distance from mirror to scale is 1 meter. The maximum angular deflection is about 14 deg. Practically all pivot instruments use pointers. Full-scale deflection corresponds to approximately 90 deg in the conventional types, which may be expanded to 120 deg in special cases.

11. Switchboard instruments and some panel instruments with deflection angles of 240 to 300 deg are becoming popular for power boards. Such instruments incorporate, for the moving-coil types, magnetic systems of the single-air-gap type; special vane arrangements are used for the moving-iron types. The large angle and resulting long scale lead to ease of reading. However, the necessity for compromise in the mechanism design to attain this wide angle of deflection usually limits the accuracy to 1 per cent. The requirement for increased power in the instrument, a limited frequency coverage in a-c instruments, and the necessity for using iron cores in the electrodynamometer types all combine to limit the use of this general form of instrument, and it is infrequently used in the electronic laboratory and in electronic gear.

12. Period and Damping. The moving element of every deflection instrument provided with a restoring torque proportional to the angular deflection is in effect a torsional pendulum. As such it has a moment of inertia P, a period T, and a damping factor. If the damping factor is low, the instrument will oscillate several times about its position of rest, each oscillation being less than the preceding one in accordance with the decrement of the system. For most rapid indication it is desirable that the instrument be not quite aperiodic or deadbeat but rather that it overswing from 3 to 5 per cent. (For a complete discussion of this see Forest K. Harris, "Electrical Measurements," Chap. 3, Equation of Motion, *et seq.*, p. 46, Wiley, 1952.)

Normal ammeters and voltmeters may be expected to have a period of the order of 1 to 2 sec. The smaller instruments, if equipped with magnets for very high gap densities and extremely light moving elements, may have a period as short as 0.2 sec (Weston high-speed power-level indicators). Instruments of ultrahigh sensitivity, where very little energy is available, may have a period as high as 5 sec. Sensitive suspension galvanometers may have a period as long as 12 sec.

The period of an instrument is important because the time necessary for any deflection instrument to attain a new position when its deflecting force is altered cannot be less than its period. High-speed indication in indicating instruments is very desirable, particularly when the phenomena being observed are rapidly changing, as in the monitoring of voice-frequency circuits; instruments with a long period will integrate the energy while high-speed instruments will give indications of peaks.

Air friction and residual pivot friction are not sufficient to prevent the moving coil from oscillating back and forth about its equilibrium position when a deflecting force is applied, and additional *damping* means must usually be provided. The amount of damping is measured by the rate at which the amplitude of the oscillations decreases. The ratio of any two successive swings is constant. The Napierian or hyperbolic logarithm of this ratio is called the *logarithmic decrement* of the instrument. The smallest amount of damping which will cause the coil to come to rest with no oscillation whatever is called the *critical damping*, and the coil is said to be critically damped. Increasing the damping beyond this point increases the time necessary for the coil to come to rest and produces overdamping. The shortest time in which the coil can come within a given small distance of its position of rest occurs when the coil is slightly underdamped. It has a value of about 1.5 times the period of the coil. The extra damping necessary to critically damp a coil is usually obtained magnetically from the motion of the coil in the field of the permanent magnet, which sets up counter electromotive forces. The amount of damping produced by the current in the coil depends upon the total resistance of the coil and connected circuit. That resistance which produces critical damping is called the *critical damping resistance* (CDR). A galvanometer is usually so designed that its critical damping resistance is at least five times its coil resistance so that it may be shunted for critical damping without losing much sensitivity. All but the most sensitive pivot instruments are damped on open circuit by the current set up in the metal winding form, and the resistance of the connected circuit has little effect on the damping.

13. The current sensitivity of any galvanometer varies directly as the number of turns on its moving coil and as the square of its period. For a given winding space on the coil, its resistance varies as the square of the number of turns, assuming that the portion of the winding space occupied by insulation remains constant. The deflection is proportional to the current and to the square root of the resistance, *i.e.*, to the square root of the power dissipated in the coil.

In the selection of galvanometers it should be noted that in general those of high sensitivity will also be slow in action, and in general the natural period and critical damping resistance for a galvanometer as listed by the several makers should be considered as carefully as the sensitivity. Further, galvanometers of highest sensitivity require great care in leveling; they are responsive to minor vibrations and in many installations may require special supports.

14. Galvanometer Supports. Where vibration in a building is a factor, the Julius suspension may be used, a somewhat complex system of weights supported by springs with oil-damping vessels. A simpler method although not so perfect is to rest a 200-lb block (of concrete) on an air cushion; this will absorb all vibration usually encountered in factories, at least for galvanometers of moderate sensitivity. Galvanometers with a single suspension have the greatest sensitivity, those with a taut suspension less, and those with double pivots least. For the most sensitive type of galvanometer, increasing the period from 5 to 40 sec allows the power to be decreased from 11 to 0.005 $\mu\mu$w. The minimum current sensitivity is 10^{-11} amp per mm. The smallest current sensitivity for a taut suspension is 10^{-8} amp per mm, and for a double-pivot pointer instrument, 5×10^{-8} amp per scale division.

15. Differential Galvanometer. Galvanometers of the suspended type are used mainly as null indicators for d-c bridges and potentiometers and as deflection instruments in comparison methods. In the latter case a *differential galvanometer* is sometimes used. This is a galvanometer having two separate insulated windings on the

Table 1. Characteristics of D-c Galvanometers

Make	Cat. No.	Sensitivity*		T, sec	Resistance, ohms	
		I, μa	E, μv		System	CDR
Suspended-coil type with mirror						
Leeds & Northrup	2285a	0.0025	0.1	7.5	17	25
	2285b	0.0035	0.2	5	16	40
	2285f	0.00004	2.8	20	800	70,000
	2285h	0.008	24	1.5	515	2,500
	2290	0.00001	1	40	800	100,000
	2284a	0.008	0.5	1.5	21	40
	2284b	0.002	0.05	7	16	10
	2284e	0.005	7.5	1.5	300	1,200
Suspended-coil type with self-contained lamp and scale						
Leeds & Northrup	2430a	0.0075	0.5	3	17	50
	2430c	0.005	2.0	2.5	25	400
	2430d	0.0005	13	3	550	25,000
	2420b	0.24	25	3	26	80
	2420c	0.025	400	3	1,000	15,000
Taut-suspension type with pointer and scale						
Leeds & Northrup	2310a	2	64	2.5	12	20
	2310c	0.25	500	3	250	1,800
	2320b	4	280	3	20	50
	2320d	0.5	1,700	3	1,000	2,400
Double-pivot type with pointer and scale						
Weston	440	0.05	3,000	2.5	4,200	55,000
	440	0.1	950	2.5	1,040	8,400
	440	0.25	300	2.7	150	1,000
	440	0.5	100	2.5	50	150
	440	2.2	30	2.3	4	10

* Sensitivity values of current I and voltage E are for a scale deflection of 1 mm at a scale distance of 1 m for suspended-coil galvanometers with mirrors; for those having self-contained scales the values given are for a deflection of the smallest division, usually 1 mm. Voltage values are on the basis of a circuit including the galvanometer and its external critical damping resistance, CDR, in series. System resistance is at the galvanometer terminals and includes both coil and suspension or spring resistance.

suspended coil. They have equal numbers of turns and are so connected that, when equal currents flow through the two coils, no deflection is produced.

16. Universal Shunt. The sensitivity of a galvanometer is most easily reduced by shunting, and, since it is desirable to keep the galvanometer critically damped, the Ayrton-Mather universal shunt shown in Fig. 3 is most convenient. This arrangement is also used in multiple-range ammeters and milliammeters and is frequently known as a "series shunt." The total resistance of the shunt is made approximately equal to the critical damping resistance of the galvanometer or indicating instrument with which it is used.

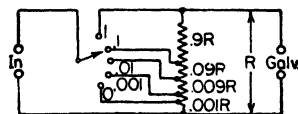



Fig. 3. Ayrton-Mather universal shunt.

17. D-c Ammeters and Voltmeters. Pointer-type instruments of the pivot type are used as ammeters and voltmeters of all ranges and as the indicating portions of thermocouple, rectifier, and various vacuum-tube instruments. The minimum range

of the ammeters extends from 5 μa to an upper limit determined only by the size of shunt desired, commercial shunts having been made to 50,000 amp. Above 15 to 30 ma the movements are shunted, in which case the copper or aluminum winding of the moving coil must have sufficient manganin swamping resistance in series with it to give a good temperature coefficient when shunted by the manganin resistance.

Alternatively, a low-resistance negative-temperature-coefficient resistor may be used in series with the moving coil, or as in the case of the more precise laboratory instruments, a shunt-series network of manganin and copper may be used for exact compensation over a selected ambient-temperature range.

Voltmeters may be made with a full-scale range from 1 mv to as high as series resistance can be arranged to care for the requirements. Instruments are made with self-contained series resistance up to a few hundred volts; higher ranges usually require an external resistor with the instrument placed in the grounded or low-potential side of the circuit for the sake of safety and to reduce electrostatic effects on the moving system.

Voltmeter sensitivity is very commonly 1 ma for full scale, although laboratory standards of certain types may take as much as 100 ma in the interest of a balanced design for high precision. Conversely, for analyzing electronic circuits where the networks are of high resistance, a full-scale value of 50 μa is quite common, giving 20,000 ohms per volt.

While series resistors for voltmeters have classically been of the wire-wound type to attain the required accuracy and stability, the advent of the pyrolytic carbon and borocarbon fixed resistors of superior characteristics has changed the picture and many panel instruments and test sets are now supplied with these types. If held to well within their rated dissipation limits for full-scale deflection they are usually satisfactory. Similar resistors incorporating thin metal films are also being used to a limited degree.

For ranges of over 1,000 volts tubular resistors are widely used incorporating wire-wound sections, each electrostatically shielded, assembled in insulating tubes and filled with inert wax.

Table 2. Voltmeters

Full scale, volts	Full scale, approximate ohms per volt	Approximate fixed resistance, ohms	Approximate 60-cycle impedance of rectifier and d-c instrument at full scale, ohms
150	1,000	149,400	600
50	1,000	49,400	600
10	1,000	9,400	600
4	1,000	3,400	600
3	2,000	4,860	1,140
2	2,000	2,800	1,140
1.5	2,000	1,860	1,140
1	5,000	3,050	1,950
0.5	5,000	550	1,950

18. Power Sensitivity. In general, pointer-type indicating instruments can be made to give full-scale deflection on as little as 0.1 μw, although for a rugged instrument from 1 to 5 μw is required. Moving-element resistances may be made from about 1 to 10,000 ohms. Low-resistance elements are limited by the spring or suspension resistance, which becomes a very appreciable part of the total, reducing the energy available for torque; high-resistance elements are limited by the available wire, and many are wound of enameled copper wire 0.001 in. in diameter.

19. Optimum Circuit Conditions. As in the output circuits of vacuum-tube amplifiers, the resistance of the instrument or galvanometer should be matched to the circuit in which it is placed for maximum energy transfer, and this is particularly

important where the energy is limited. On the other hand, this will frequently result in overdamping galvanometers of ultrahigh sensitivity, and a compromise must usually be made between speed of response and sensitivity requirements. It should be noted, however, that this matching is not of vital importance since the loss by an error in match by as much as 20 per cent is very small.

Table 3. Characteristics of A-c Galvanometers as Listed by Leeds & Northrup

Cat. No.	Sensitivity*		T, sec	Resistance, ohms	% resonance range	
	I, μa	E, μv			Current	Emf
Vibration galvanometer, with mirror, for 60-cps use						
2350a	0.025	500.0	...	800	0.3	1 5
2350b	5.0	2.0	...	0.35	0.3	0.7
With pointer and scale for 60-cps use						
2370a	5.0	.....	2.6	20	Separately excited field requires 40 watts at 115 volts, 60 cps	
2370b	1.0	.....	2.6	250		
2370c	0.5	.....	2.6	1,000		

* Sensitivity values of current I and voltage E are for a scale deflection of 1 mm at a scale distance of 1 m for the vibration galvanometers; for the pointer-and-scale type the current values given are per 1-mm division.

20. Moving-coil Vibration Galvanometers. When an alternating voltage is applied to the coil of a permanent magnet galvanometer, the coil will follow the alternations of the current if the frequency is of the same order as that defined by its period. Maximum amplitude of vibration will occur at the natural frequency of the coil. The relation between amplitude and frequency is similar to the resonance curve of an electrical circuit. The ratio of the maximum amplitude at its natural frequency to the amplitude for an equal d-c voltage is between 25 and 150. The period of the ordinary d-c galvanometer is never less than 1 sec, while the frequencies at which measurements are made are rarely less than 30 cps. The upper limit for a taut single suspension is around 300 cps. This limit may be raised to 1,000 by the use of a taut bifilar suspension. Electrical characteristics of commercial vibration galvanometers are given in Table 3. At 60 cycles their sensitivity is equal to that of a good d-c galvanometer. A resonance curve when tuned to a frequency of 100 cps is shown in Fig. 4.

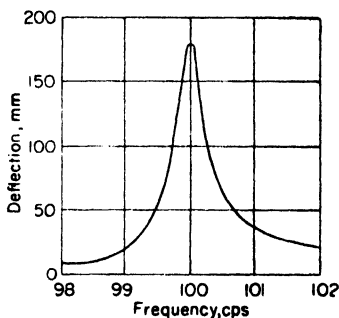



Fig. 4. Resonance curve of vibration galvanometer.

The natural frequency may be raised still further by eliminating the coil entirely and using the single-turn loop formed by the bifilar suspension. The mirror is then placed at the center of the taut wires. The general method of construction is shown in Fig. 5. By this means a natural frequency of 12 kc may be obtained. The sensitivity decreases inversely as the first power of the frequency. On this account it is as sensitive at 10 kc as the bifilar-coil galvanometer was at 1 kc. In comparison with other null detectors at these frequencies, its sensitivity is so low that it is not much used in this form.

The Einthoven string galvanometer uses the simplest possible moving system for a galvanometer. A single conducting string moves in the narrow air gap of the magnetic system, which may be a permanent magnet or an electromagnet depending on the sensitivity desired. Its motion is observed through a microscope or by its shadow thrown on a screen from a point light source. The Einthoven string galvanometer built by the Cambridge Instrument Company uses a gilt glass fiber of 0.002, 0.003, or 0.005 mm diameter with a resistance of 6,000, 4,000, or 1,400 ohms, has a period of 0.0036 to 0.01 sec, and at 1 m will deliver deflections of 2 to 200 mm per μa. The string galvanometer may also be used as an oscillograph. The shadow of the string is observed on a translucent screen as reflected from a revolving mirror. The motion of the string may also be photographed on film or bromide paper. The usual paper speed is 10 in. per sec, but this may be increased to a maximum of 100 in. per sec. At this latter speed, phenomena lasting 1 msec appear 0.1 in. long.

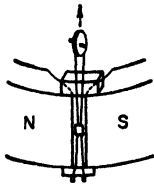



FIG. 5. Bifilar suspension.

21. Moving-coil A-c Instruments. If a steady deflection is desired with a.c., the magnetic field must change in direction with the current in the coil and must have the same phase. This requires that the field be electromagnetic. In the case of galvanometers and particularly null indicators, a field of laminated iron may be used, excited at the same frequency as the moving coil. When used as a null indicator in a bridge network, the field is connected across the same supply as the bridge, while the moving coil is connected to the detector terminals. Since the current through the field and the flux produced will be nearly 90 deg out of phase with the voltage applied to the bridge, the galvanometer will be most sensitive to the reactance balance and will be little affected by the resistance balance. These conditions may be equalized or reversed by the introduction of resistance in series with the field, or reactance in series with the bridge, to make the field current and bridge current differ in phase by 45 deg or be in phase. The phase selectivity of the a-c galvanometer may be of advantage in certain special cases, but in general it is a considerable disadvantage. The electrostatic field of the main field winding exerts a considerable force on the moving coil so that it must be carefully shielded. The sensitivity of the a-c galvanometer is very high and it compares favorably with the best d-c galvanometers.

22. Electrodynamometer. When the iron core is omitted from the field winding, the moving coil and field coil may be connected in series. The deflection is then proportional to the square of the current flowing in the windings, and the instrument is called an *electrodynamometer*. Instruments of this type read the same on a.c. and d.c. and are suitable as transfer instruments, provided certain precautions are taken. Protection from external magnetic fields is most important. This is usually accomplished in pivot-type instruments by shielding with soft iron. It may also be effected by making the instrument *astatic*. When a.c. is used, an error is introduced if the distribution of current in the coils is affected by eddy currents in the conductors themselves—the so-called *skin effect*—or by capacitance between windings. The former effect is minimized by the use of conductors with insulated strands—so-called *litzendraht*—the latter by careful spacing and by electrostatic shielding.

Electrodynamometers may be used as galvanometers, ammeters, voltmeters, and wattmeters. Their sensitivity as galvanometers is so low compared with vibration galvanometers and other meters that they are now rarely used. As ammeters, voltmeters, and wattmeters, they are the standard instruments for use at commercial frequencies. In general the sensitivity of a-c instruments is of the order of 1/1,000 of that of d-c instruments, this being due to the difference in field intensity of the electromagnetic field as compared with that which can be obtained from a permanent magnet. Electrodynamometer instruments of the highest precision will take from 1 to 3 watts full scale, the total energy varying with the square of the deflection. Suspension-type electrodynamometers may have sensitivities 100 times as great.

Electrodynamometer ammeters have their fields and moving coils in series up to

several hundred milliamperes above which the moving element is shunted across a resistor in series with the fixed coils. Above 50 amp or so, current transformers are used, and these are now available with special alloy cores which will give accuracies of the order of ⅒ of 1 per cent. Note that low-range voltmeters have very low resistance in order to get the required energy; dynamometer voltmeters with full-scale values of 2 volts may draw as much as 0.5 amp. High voltages above 1,000 volts are measured with potential transformers.

Table 4. Characteristics of Typical A-c Instruments

Make, type	Model	Accuracy, %	*E*, volts	*I*, amp	*R*, ohms	*W*, watts
			Electrodynamometer type			
Weston portable....	341	0.25	30/15	......	200/100	4.50/2.25
	341	0.25	300/150	......	6,680/3,340	13.50/6.75
	370	0.25	....	10/5	0.072/0.18	7.20/4.50
			Moving-iron type			
Weston portable....	904	0.5	300/150	......	13,600/6,800	6.60/3.30
	904	0.5	....	10/5	0.0063/0.022	0.63/0.55
	433	0.75	30/15	......	336/168	2.68/1.34
	433	0.75	300/150	......	22,000/11,000	4.10/2.05
	433	0.75	....	5	0.02	0.50
Weston panel.......	517-476	2.0	15		212	1.07
	517-576	2.0	150		16,000	1.41
	517-476	2.0	....	5	0.010	0.25
			Thermocouple type			
Cambridge.........	....	....	....	0.010	30	0.003
				0.10	4	0.04
				1.00	0.2	0.2
Weston portable....	622	0.5	....	0.002	800	0.0032
	622	0.5	....	0.010	25	0.0025
Weston panel.......	507-425	2.0	....	0.12	5.00	0.072
	507-425	2.0	....	1.0	0.25	0.250
	507-425	2.0	....	10.0	0.02	2.000

Values of voltage *E*, current *I*, and power *W* are for full-scale deflection.

23. Electrodynamometer Wattmeters. When the field coils are connected in series with the load, and the moving coil of fine wire in series with a high resistance is placed across the load, the deflection is proportional to the product of the load voltage and the in-phase value of the current, or $EI \cos \theta$, or watts. For polyphase circuits a multiplicity of similar elements may be arranged on a single shaft, the most usual variety being the two-element instrument on three-phase circuits. Such an instrument gives true power without relation to phase angle.

24. Moving-iron Instruments. Galvanometers may be constructed with a stationary coil and a moving-iron vane or magnet. The moving system consists of small permanent magnets placed at the center of the coil at right angles to the axis of suspension. To avoid the effect of outside magnetic fields, the system is duplicated with the magnets pointing in the opposite direction to make it astatic, and the whole galvanometer is surrounded by multiple soft-iron shields. Its sensitivity is nearly equaled by the best moving-coil galvanometers so that it is very little used.

Soft iron may also be used in the moving element, either alone or in conjunction with a fixed piece of soft iron, both of which are magnetized by the fixed coil.

Soft-iron meters are much used as a-c ammeters and voltmeters in a wide variety of ranges and sizes. Electrical characteristics are given in Table 4. The range of the ammeters is from 20 ma to 500 amp. The upper limit is ten times that of dynamometer-type meters, because the current coil is fixed. Currents up to 5,000 amp are measured by the use of current transformers. The range of the voltmeters is from 1 to 750 volts. The resistances are such as to give from 3 to 200 ohms per volt, the values increasing with the voltage. Higher voltages are measured by the use of either multipliers or potential transformers. Moving-iron instruments used on d.c. may show gross errors due to residual magnetism in the vanes; the errors, up to 5 per cent, depend on previous peak values. They indicate true rms values on a.c., provided that on *peaked* wave forms the iron vanes have not saturated, and this is usually the case if the *peak* value of current or voltage is not over four times the full-scale rms value. Wave-form errors are discussed below.

25. Frequency Compensation. Moving-iron ammeters, if calibrated at 60 cps, usually show minor errors at higher frequencies, mainly due to eddy currents in the vanes and adjacent masses of metal. The magnitude of the error at 400 cps is generally small, however, and with the poorest designs of 90-deg deflection instruments will rarely be over 2 per cent. At 1,000 cps the best ammeters have a frequency error of not over 1 per cent. Long-scale (*i.e.*, 240 deg) ammeters may have gross errors (up to 50 per cent) at higher power frequencies and, if used outside their rated span, should be specifically checked at the frequency in question. The best moving-iron ammeters may be adjusted for use at a particular frequency to as high as 10 kc.

An uncompensated moving-iron voltmeter has a much larger error at the higher power frequencies. Basically a milliammeter in series with a calibrated resistance, the coil reactance and the meter impedance rise with frequency so that the indication drops rapidly with the increase in frequency and may be 10 to 50 per cent low at 1,000 cps. However, by shunting the resistance with a capacitor the coil current can be maintained substantially uniform at constant voltage and varying frequency. And, further, a slight overcompensation of the impedance will cancel the error due to eddy currents to give a reasonably flat voltage response up to several thousand cycles. The circuit is shown in Fig. 6.[1] An uncompensated voltmeter showing an error of 12 per cent at 1,000 cps can be compensated to have an error due to frequency of less than ¼ of 1 per cent from 60 to 1,100 cps. If greater coverage is desired the errors at intermediate points are greater so that, if the previously mentioned voltmeter were compensated at 1,500 cps, the maximum error would be 1 per cent at about 1,100 cps. Electrodynamometer voltmeters can also be compensated in a similar manner.

FIG. 6. Circuit diagram of compensated voltmeter.

In general the sensitivity of pointer-type indicating instruments using the moving-iron principle is from 0.1 to 1 watt full scale. Instruments using short vanes, usually of the arcuate type, take about 1 watt full scale. Instruments with long radial vanes are more sensitive with a minimum of 0.1 watt full scale but in general are more sensitive to external fields and must be well shielded and kept away from strong external fields. Moving-iron instruments are widely used on power circuits and are generally available in all sizes from the small 2-in. instruments up to the larger switchboard types.

HIGH-FREQUENCY CURRENT METERS

Because of the large reactance of any electromagnetic device at the higher frequencies, it is difficult to design such instruments for frequencies over 3,000 cps; while

[1] Miller, John H., Frequency Compensation of A-C Instruments, *Trans. AIEE*, **70** (6), 217, June, 1951.

occasionally used at 10 kc they may consume as much as 30 va for full-scale deflection. Above 10 kc some converting device is used, usually to produce a small amount of d.c. which is a function of the high-frequency current in question.

26. The Thermoammeter. If high-frequency current heats a short piece of fine platinum wire, and the temperature of the middle of the wire is measured with a thermocouple connected to a sensitive d-c millivoltmeter, the millivoltmeter scale can be calibrated in terms of the current in the heater. Such a simple assembly, however, does not compensate for variations in temperature of the terminals or for ambient-temperature variations.

The Weston thermal ammeter as developed by W. N. Goodwin, Jr., is as shown in Fig. 8. The heater is a wire or tube of platinum alloy of very short length whereby most of the heat is conducted to the terminals, thus wiping out largely the effect of convection currents of air. The temperature of the heated member may be represented as a parabola in its gradient from center to each terminal lug, and it is this temperature difference or gradient from the center of the heater to its end which is measured by the thermocouple. The couple proper consists of a pair of wires, usually of constantan and a

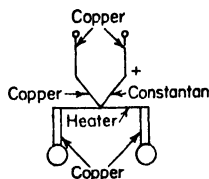



FIG. 7. Thermocouple meter.

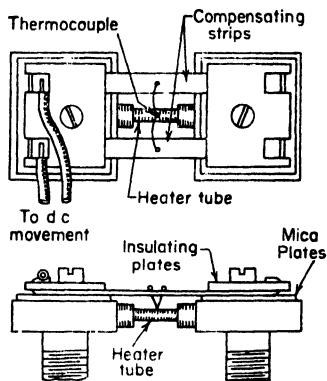



FIG. 8. Compensated high-frequency thermocouple and heating element.

platinum alloy, permanently welded to the center of the heater at the junction end, with the effective cold ends soldered to a pair of copper strips which are thermally connected to, but electrically insulated from, the terminal lugs. Their heat capacity is such that the difference in temperature between the center of the heated member and the center of the two copper compensating strips is always the same as from the center of the heated member to the terminal lug, regardless of ambient-temperature changes or general rise in temperature of the surroundings due to heating of the lugs themselves or temperature rise due to the total heat generated. The thermoelectric voltage is, therefore, strictly proportional to the temperature difference between the center and ends of the heated member which in turn is proportional to the square of the current causing this temperature rise, and a d-c instrument connected to the couple may be calibrated in terms of this current.

27. Thermoelements. These may be designed to give suitable indication on instruments of commercial types from 200 ma up to whatever may be required. Solid round wires may be used for the heated member up to about 2 amp, but for higher currents and at the higher frequencies skin-effect phenomena cause the readings to be too high. For higher ranges, therefore, the heated member should preferably take the form of a thin-walled tube of such dimensions that the frequency error will not be excessive. For a frequency error of 1 per cent at 50 Mc, and for a platinum alloy tube having a resistivity of 170 ohms per cir mil-ft, the wall thickness is about 0.001 in.; at 100 Mc the readings will be about 5 per cent high. Since a 1-mil wall represents a practical minimum thickness of the heated tubular element, thermoammeters are not useful for even moderate accuracy at frequencies much higher than 150 Mc; as a practical matter above 150 Mc standing waves of even minor magnitude preclude accurate measurement of current values as such, and the intro-

duction of a line discontinuity by adding the thermoelement may also cause gross errors. Current measurements in the usual sense are, therefore, rarely made at above 150 Mc. It should be noted, however, that as long as the pointer of a thermoammeter does not deflect beyond full scale, no damage will occur to the thermoelement itself at any high frequency, and qualitative measurements may be made on frequencies much higher than 150 Mc.

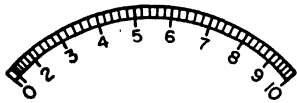
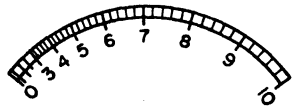



FIG. 9. (*Above*) Standard scale using a conventional d-c movement. (*Below*) Linear expanded scale using specially shaped pole pieces.

While standard instruments have a square-law scale as the result of the I^2R production of heat, instruments are available in which the upper four-fifths of the scale is approximately linear through the use of special d-c indicating mechanisms having nonlinear air gaps whereby the d-c sensitivity is progressively lower as the pointer moves up the scale. By a proper combination of such specially shaped pole pieces a nearly linear scale may be produced. (See Fig. 9.)

Instruments having the linear expanded scale are useful in small broadcast transmitters licensed for a lower power at night than during the day; sufficiently accurate readings of the high and low values of antenna current may be had on the same instrument to be satisfactory, and instruments of this type are listed as complying with FCC rule No. 143.

For low ranges so-called *bridge-type* thermoelements are used, as shown in Fig. 10, whereby a number of couples are arranged in series-parallel to give a higher thermal emf. The impedance of these couples is higher than for a single couple, and for the common current-squared galvanometer the effective resistance is 4.5 ohms. The indicating instrument for the standard single-unit thermoelement has a sensitivity of 12 mv and a resistance of about 5 ohms.

FIG. 10. Galvanometer or bridge-type couple.

28. Vacuum Thermoelements. For still higher sensitivities the assembly may be placed *in vacuo*. Such thermoelements show no increase in sensitivity until the vacuum is better than 0.01 mm. of mercury; as the vacuum is increased, the sensitivity of the low ranges with very fine heaters increases markedly since there is no cooling by convection currents. Characteristics of a line of vacuum thermoelements are listed in Table 5, as made by Weston, and are typical. Metallic heaters are used on the higher ranges, with a carbon filament for the heater of the 1.5- and 2.0-ma ratings.

Both contact and insulated types are made and listed. In general the contact type is more efficient, and the output resistance is lower. In the insulated type a tiny ceramic bead connects the couple junction and the heater, thus insulating the two while allowing heat to flow from the heater to the junction. The equivalent capacitance across the bead is less than 1 $\mu\mu$f. Because less heat flows across the bead, the couple wires must be smaller and the output resistance is higher as listed.

Vacuum thermoelements are quite rapid in response. If a given current is suddenly sent through the heater and maintained, the output millivolts will increase in a manner departing only slightly from an exponential curve, and we may thus define the "time constant" as the time for the output to rise to 63 per cent of its steady-state value.[1] For all except the carbon-filament ranges, the time constant so defined

[1] Harris, Forest K., Letter to the Editor, Thermal Converter Time Constants, *Elec. Eng.*, **75** (7), 674, July, 1956.

Table 5. Characteristics of Vacuum Thermoelements

Range rated, ma*	Heater ohms ± 10%	Approx heater drop at rated current, mv	Couple ohms ±10%		Max safe heater current, ma	Frequency for 2% eror, Mc‡
			Contact type	Insulated type		
1.5	1,365	2,000	7.5	10	3.2	300
2	750	1,500	6	10	5.0	300
2.5	475	1,200	6	10	6.2	300
4	95	380	9	10	9.0	300
5	90	450	6	10	10.0	300
7.5	40†	300	6	10	16	300
10	27†	270	6	10	25	300
15	13	195	6	10	40	300
20	8.4	170	6	10	50	300
25	7.0	175	6	10	62	300
30	5.8	175	3	5	75	300
37.5	4.6	170	3	5	85	300
50	3.3	165	3	5	115	300
75	1.36	100	3	5	170	200
100	1.03	100	3	5	220	200
150	0.66	100	3	5	320	150
200	0.44	90	3	5	420	75
250	0.39	100	3	5	510	75
300	0.33	100	3	5	610	50
400	0.25	100	3	5	800	35
500	0.20	100	3	5	1,000	35

* Rated heater current will produce 5 to 6 mv d.c. at couple terminals.
† Subject to ±15%.
‡ Error due to skin effect only.

is between 0.1 and 0.2 sec for the contact types (including higher ranges in air as previously described); the insulated types will run 0.2 to 0.3 sec because of the time lag of the bead.

These thermoelements are to be used with 10-ohm, 200-μa instruments, and are used either external to the instrument or self-contained in the instrument case. In some instances they are associated directly with the coaxial line to simplify the r-f system, with leads to the instrument. Frequency errors will vary with range; up to 150 Mc such errors are usually under a few per cent. Care must be taken to avoid r-f currents in the couple which will heat the junction independently of the heater current; chokes in the instrument leads and capacitance to ground may be required to maintain this condition.

FIG. 11. Physical dimensions of thermoelements listed in Table 5.

29. Thermoammeters in general are calibrated on commercial frequency a.c.; if used on d.c., the mean of reversed readings should be taken to make certain that any d-c drop in the heater picked up by the couple is canceled out. They may be obtained with a separate thermoelement for use in indicating at a distance as where a thermoelement is placed in the antenna of a transmitting station and the leads brought back to an instrument in the transmitting building. The thermoelement should be placed in the h-f circuit at a point close to ground potential to reduce circulating r-f currents in the leads to the instrument. If this cannot be done, the thermoelement,

of low range, around ½ amp, is connected to a loop of wire that is inductively coupled to a loop in the main antenna circuit. The thermoelement circuit may then be grounded. The instrument is scaled to read the total antenna current, and the final calibration is made by adjusting the inductive coupling between the two loops until the remote-reading instrument indicates the same value as an instrument placed directly in the antenna itself. Note that FCC rules require an instrument in the main antenna circuit which may be used for this purpose but which under normal operating conditions is kept short-circuited to prevent damage due to lightning. The switch is opened when the instrument is read, and the remote indicator, usually located on the transmitter panel, is used for normal operation.

The ratio of the power available to operate the indicating meter to that put into the heater is about 1 to 2,000 for the most efficient couples; hence a very sensitive d-c instrument is required for low r-f energies.

30. Thermocouple voltmeters are constructed by using one of the more sensitive couples with sufficient series resistance to give the desired voltage range. Their range is from 0.3 to 150 volts with resistances of 125 ohms per volt above 1 volt, and 500 ohms per volt above 10 volts, if desired. Their frequency range is determined by that of the series resistance. The small resistance spools which must be used in meters with self-contained resistors change in resistance rapidly with frequency so that their frequency limit is 3 kc. Frequencies of 1 Mc may be attained with an error of 1 per cent with special h-f resistors.

Since the emf produced by the thermocouple is proportional to the power input and hence to the square of the current, this meter will read correctly on both d.c. and a.c. and may, therefore, be used as a transfer instrument. It is necessary, however, to take the average of the readings for both directions when using d.c.

RECTIFIER METERS

31. Rectifier Meters. The full-wave preferred type of rectifier meter includes a bridge rectifier with four rectifier disks in circuit with a d-c meter as shown in Fig. 12; the figure also shows a series resistance R making this a rectifier voltmeter. With perfect rectifiers having zero forward resistance and infinite back resistance, if 1.0 ma a.c. is passed into this network and is measured with an rms or effective-value instrument, the meter receives a pulsating d.c. having a value of 0.9 ma since the d-c meter indicates the average value. The reciprocal of this ratio, 1.11, is the *form factor* of a sine wave, equal to $\pi/(2\sqrt{2})$, and this represents the ratio of the rms value of the sine-wave input into an ideal bridge rectifier, to the average value of its d-c output.

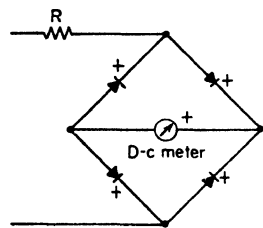



FIG. 12. Copper-oxide rectifier bridge.

Rectifier meters as milliammeters, or with appropriate series resistance as voltmeters, are usually calibrated on sine-wave a.c., or voltage, in terms of the rms value. On d.c. or on square-wave a.c. they will therefore read 11 per cent high. And on other wave shapes they will read in terms of the form factor of that wave shape.

32. Rectifiers Used in Rectifier Meters. In spite of the availability of excellent silicon and germanium rectifiers, copper-oxide rectifiers are the preferred type used in most rectifier meters. Since the drop across the instrument mechanism itself is of the order of a few hundred millivolts and the resistance of the instrument moving coil is low, the relatively low forward resistance and, more particularly, the low voltage drop across copper oxide gives the best impedance match. This factor is particularly important in rectifier milliammeters since a milliammeter is placed in series with a circuit in which, presumably, it will have little or no effect and maintaining a low resistance is quite important. Likewise for measuring low a-c voltages a low resistance of the rectifier meter is desirable so that the maximum amount of series resistance may be used to swamp out to the greatest degree the effect of temperature on the

measuring system. On the other hand, for rectifier-type voltmeters having ranges of over 50 volts, or for rectifier-type milliammeters to be used in high-impedance circuits, the silicon diode becomes the preferred rectifier. Thus, while the standard rectifier meters continue to use copper-oxide rectifiers, rectifier meters may be obtained from most makers on special order including silicon diodes where the added accuracy and perfection of rectification are worth the considerable extra expense involved.

Other rectifying systems such as those using selenium or copper-sulfide rectifiers are not used because of the high drop at low currents.

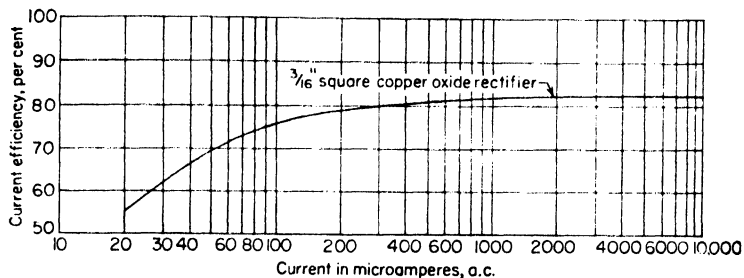



FIG. 13. Current-efficiency characteristic.

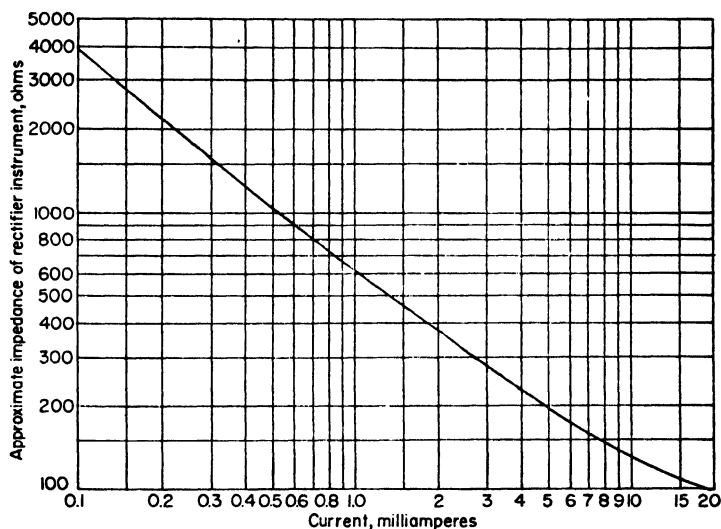



FIG. 14. Impedance characteristic of 20-ma rectifier instrument at 60 cycles.

33. Commercial Rectifier Instruments. These usually contain the full-wave rectifier described above, and this applies as well to power-level indicators and VU meters discussed later. However, many multiple-range instruments for radio servicing use a half bridge or some modification of a half-wave system as an economy; the lower efficiency can usually be tolerated since practice in these assemblies generally dictates a d-c voltmeter sensitivity of 50 μa full scale, to give d-c voltage ranges at 20,000 ohms per volt, whereas the a-c ranges are at 1,000 ohms per volt.

Because there is a small back current in the copper-oxide rectifier the current efficiency, defined as the ratio of the average d-c output to the rms a-c input, is less than the 90 per cent theoretical maximum efficiency and Fig. 13 shows the current efficiency of a typical copper-oxide rectifier for instrument use. However, since

rectifier meters are adjusted and calibrated as an assembly, this slightly lower efficiency is not material and is calibrated into the meter as a whole.

In general the impedance of the rectifier rises with lower currents, and Fig. 14 shows typical values of effective impedance for a 20-ma rectifier instrument with current.

Temperature variations have considerable effect on both the impedance and accuracy of rectifier instruments. Figure 15 shows temperature-voltage variations for a specific group of milliammeters from which impedance can be determined. Figure 16 shows temperature-efficiency relations of this group at various current values. The point must be stressed, however, that the curvature of these characteristics varies with the several parameters of rectifier-disk size, current density, processing time, and the resistance of the instrument, and it is quite possible to modify these curves materially for special requirements. Standard instruments, by the same token,

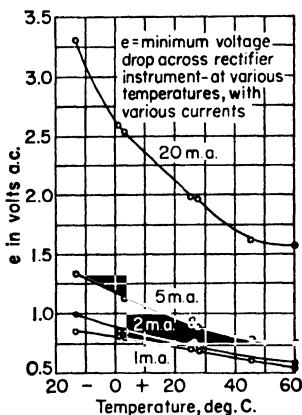



FIG. 15. Effect of ambient temperature on the voltage drop across a rectifier instrument at various currents.

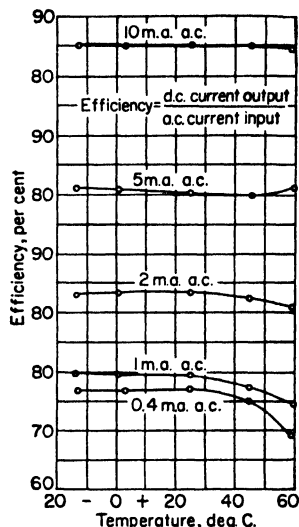



FIG. 16. Ambient-temperature efficiency relation for several typical rectifier-type milliammeters.

can hardly be represented by any particular group of curves. Rectifier instruments have been materially improved in recent years as to the flattening of the curves, and design possibilities have broadened to the point where materially improved instruments can be made for particular requirements.

Higher temperatures adversely affect the rectifying film, and rectifier instruments may become erratic at temperatures in excess of 45°C. High-temperature locations should be avoided; where the instrument becomes unduly warm, instruments with external rectifiers are sometimes used with the rectifier placed in a relatively cool location.

Frequency errors are the result of capacitance between disks. Since the disk resistance is lower at higher currents and since capacitance is a function of rectifier size, the smallest rectifier is preferred for good frequency characteristics. This in turn means a high current density with which good accuracy is obtainable somewhat above audio frequencies. With low-current density, errors may be as large as 1 per cent per 1,000 cycles.

In general, low-range voltmeters are more subject to temperature and frequency errors than high-range voltmeters. Low-range voltmeters have scales which are compressed at the lower end owing to variations of impedance with current. High-range voltmeters and milliammeters have nearly uniform scale distribution.

Table 6 gives approximate constants of commercial rectifier instruments.

Table 6. Milliammeters and Microammeters

Full Scale, Ma	Approximate 60-cycle Impedance at Full Scale*
15	100
10	130
5	190
2	370
1	600
0.5	1,140
0.2	1,950
0.1	4,200
0.05	6,300
0.02	10,000

* Individual copper-oxide rectifiers vary considerably from the average in characteristics. Impedance values given may vary ±15 per cent, and efficiency values vary ±3 per cent for the product of *one* manufacturer. Much greater variations may be expected between the products of *different* manufacturers.

34. Power-level Instruments. For monitoring power level in voice-frequency circuits the instruments are usually rectifier-type voltmeters with scales calibrated to read power on the basis of a fixed resistance load. Indications are in decibels above or below a specified zero power level; alternatively they may be in VU, amounting to decibels above or below 1 mw, and as defined in the section on the VU meter. The zero level for power-level instruments marked in decibels has been considered as 6 mw in the older versions but currently 1 mw is becoming the standard and instruments marked in decibels and using this level frequently carry the caption DBM.

The older instruments with a full-scale value of 6 db above a zero level of 6 mw in 500 or 600 ohms carry a scale as shown in Fig. 17. The instrument is basically a rectifier-type voltmeter with a resistance, measured at full scale, of 5,000 ohms and adjusted to 3.79 volts full scale and 1.90 volts at zero level. The value of 5,000 ohms was selected to minimize the loss in line energy due to the addition of the power-level indicator and also to avoid harmonics to the line due to the nonlinear shunt resistance of the instrument rectifier network.

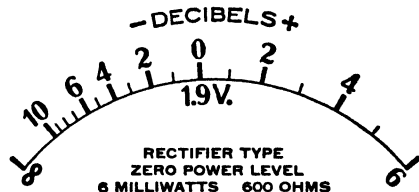



FIG. 17. Scale of typical power-level meter.

35. The VU Meter. The VU meter, which is currently the common denominator used in monitoring voice-frequency circuits, was developed in order that a standard-level indicator might be made available. It is rather completely described by Chinn, Gannett, and Morris in their classic paper.[1] The original work envisioned the instrument as always to be used across a 600-ohm source feeding a 600-ohm load. However, with the use of 150-ohm lines in more recent practice, particularly in television, an alternative arrangement for use on 150-ohm lines has been developed. And in recognition of the fact that the basic instrument may be used in conjunction with lines of any resistance, the latest edition of the standard is quite general and considers the use of the meter across lines of any stated impedance.

More specifically this standard is referred to as "IRE Standard on American Recommended Practice for Volume Measurements of Electrical Speech and Program Waves 1953" and was printed in the *Proc. IRE*, May, 1954. It is also considered as an American Standard and carries the reference C16.5-1954.

Two instrument scales have been standardized, as shown in Fig. 18. The *A scale* emphasizes the VU markings and has an inconspicuous voltage scale. The *type B scale* emphasizes the per cent voltage and has a relatively inconspicuous VU scale. This latter scale is largely used in broadcast monitoring since the voltage scale indi-

[1] Chinn, Gannett, and Morris, *Proc. IRE*, January, 1940; A New Standard Volume Indicator and Reference Level, *Bell System Tech. J.*, January, 1940.

cates in a rather direct fashion the per cent utilization of the facilities. The scales are printed on buff paper to reduce eyestrain; the narrow arc and the figures above it are in black with the heavy arc to the right, the markings above it as well as the markings below the arc in red.

The instrument mechanism, which is identical for both scales, has very definite ballistic characteristics which may be completely defined by the fact that, if a voice-frequency voltage of such amplitude as to give a steady reading of 100 on the voltage scale is suddenly applied, the pointer should reach 99 on this scale in 0.3 sec and should then overswing the 100 point by between 1 and 1.5 per cent.

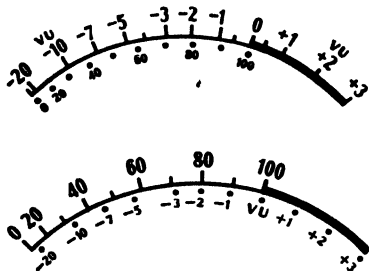



Fig. 18. (*Above*) A scale for VU meter. (*Below*) B scale for VU meter.

Since a voice-frequency channel contains many components of different frequencies, the dynamic standards listed above are a very necessary part of the specification for the instrument. The instrument is standardized on sine-wave voltage at 1,000 cps in terms of the rms value of that voltage and is adjusted to read to the 100 mark on the voltage scale with 1.22 volts applied to the standard instrument as furnished plus a 3,600-ohm external series resistance, this representing 4 db above 1 mw in 600 ohms.

With such an instrument, the readings obtained from it when voice-frequency currents are applied may then be stated as so many VU, taking into account that 4 VU must be added to the scale reading plus the number of VU lost in the attenuator placed in the network.

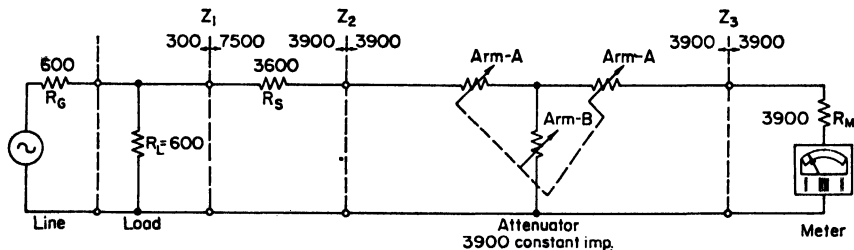



Fig. 19. Network for use with VU meter and 600-ohm line.

36. Applying the VU Meter. The VU meter is connected across a 600-ohm line feeding a 600-ohm load as shown in Fig. 19. The fundamental total resistance of the instrument is 7,500 ohms. To this is added 300 ohms representing a 600-ohm source and load in parallel, making a total of 7,800 ohms. To simplify the use of an attenuator, this is split in the center to give 3,900 ohms each siae, which will allow a simple T-pad attenuator to be inserted at this point. The instrument proper, therefore, has an internal resistance of 3,900 ohms and must be used with the separate 3,600-ohm resistor. Since the normal instrument level is +4 VU, the attenuator dial is marked 4 VU at zero attenuation, and for other true attenuation values 4 VU are added. Table 7 shows values for such attenuators.

For use on 150-ohm lines it was found impractical to change the potential and current relations in the VU meter because of the finite resistance of the self-contained rectifier, and it would also have required another type of basic instrument. Consequently another approach was investigated and it was found that the standard meter could be coupled to the 150-ohm line through an appropriate transformer which, effectively, lowered the meter impedance by a factor of 4 while maintaining

Table 7. Attenuator Network for VU Meter on a 600-ohm Line

Attenuator loss, db	Level VU*	Arm *A*, ohms	Arm *B*, ohms	Attenuator loss, db	Level VU*	Arm *A*, ohms	Arm *B*, ohms
0	+4	0	Open	15	+19	2,722	1,432
1	+5	224.3	33,801	16	+20	2,833	1,268
2	+6	447.1	16,788	17	+21	2,935	1,124
3	+7	666.9	11,070	18	+22	3,028	997.8
4	+8	882.5	8,177	19	+23	3,113	886.3
5	+9	1,093	6,415	20	+24	3,191	787.8
6	+10	1,296	5,221	21	+25	3,262	700.8
7	+11	1,492	4,352	22	+26	3,326	623.5
8	+12	1,679	3,690	23	+27	3,384	555.0
9	+13	1,857	3,116	24	+28	3,437	494.1
10	+14	2,026	2,471	25	+29	3,485	440.0
11	+15	2,185	2,388	26	+30	3,528	391.9
12	+16	2,334	2,091	27	+31	3,566	349.1
13	+17	2,473	1,838	28	+32	3,601	311.0
14	+18	2,603	1,621				

* VU = numerically equal to number of decibels above 1-mw reference level.

the appropriate power ratios. Figure 20 shows the recommended connections of the standard VU meter to a 150-ohm line. The transformer is a low-loss toroidal type

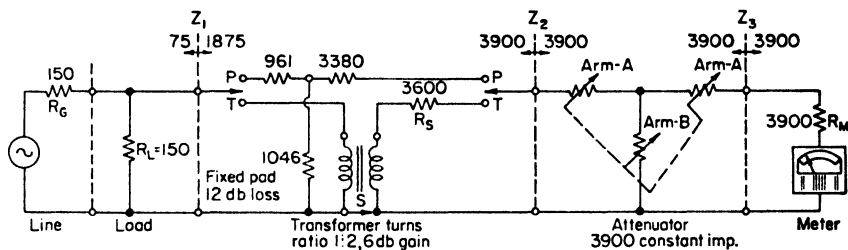



FIG. 20. Network for VU meter for use on a 150-ohm line.

such as the General Radio Type 941A, the transformer itself having its terminals connected as shown in Fig. 21. The transformer is used directly with switches *T* closed and up to any level for which the transformer is suitable. Switch *S* may be opened for a fully insulated system or omitted entirely and this circuit may be closed if a grounded system is deemed satisfactory.

FIG. 21. Terminal-connection diagram for coupling transformer associated with VU meter on a 150-ohm line.

However, since any transformer has some power limitation, and since a 12-db pad can be arranged to serve as a match between the line and load, Fig. 20 shows an arrangement where the transformer is switched out of the circuit and a fixed pad switched in at the 12-db or −16 VU point; the two line switches are transferred to *P*, and switch *S* must be closed if left open previously. Reference is made to Table 8, giving the attenuator values for this network, and it will be noted that at the 12-db position on the attenuator the transformer is taken out of the circuit and a fixed 12-db loss network is added, after which the instrument is directly connected. It is convenient to associate these transfer switches with the 12-db position on the attenuator and it might be noted that it is possible to use the resistor elements for the first 11 steps over again for the higher steps shown in the tabulation of resistance values.

Any other type of attenuator, such as a balanced H, may be used, with the matching transformer ahead of the attenuator throughout, or transferred to a fixed pad at the higher levels as may best suit the requirement of any particular level and circuit.

Table 8. Attenuator Network for VU Meter on a 150-ohm Line

Attenuator loss, db	Level VU	Switch position	Arm *A*, ohms	Arm *B*, ohms	Attenuator loss, db	Level VU	Switch position	Arm *A*, ohms	Arm *B*, ohms
0	+4	T, T	0	Open	15	+19	P, P	666.9	11,070
1	+5	T, T	224.3	33,801	16	+20	P, P	882.5	8,177
2	+6	T, T	447.1	16,788	17	+21	P, P	1,093	6,415
3	+7	T, T	666.9	11,070	18	+22	P, P	1,296	5,221
4	+8	T, T	882.5	8,177	19	+23	P, P	1,492	4,352
5	+9	T, T	1,093	6,415	20	+24	P, P	1,679	3,690
6	+10	T, T	1,296	5,221	21	+25	P, P	1,857	3,166
7	+11	T, T	1,492	4,352	22	+26	P, P	2,026	2,741
8	+12	T, T	1,679	3,690	23	+27	P, P	2,185	2,388
9	+13	T, T	1,857	3,166	24	+28	P, P	2,334	2,091
10	+14	T, T	2,026	2,741	25	+29	P, P	2,473	1,838
11	+15	T, T	2,185	2,388	26	+30	P, P	2,603	1,621
12	+16	P, P	0	Open	27	+31	P, P	2,722	1,432
13	+17	P, P	224.3	33,801	28	+32	P, P	2,833	1,268
14	+18	P, P	447.1	16,788					

Table 9. Useful Technical Decibel Data

Power level, db	Power ratio to 0 db. Also power, mw, when 0 level = 1 mw	Voltage ratio to 0 db	Voltage—based on 1 mw in 600 ohms = zero level	Power level, db	Power ratio to 0 db. Also power, mw, when 0 level = 1 mw	Voltage ratio to 0 db	Voltage—based on 1 mw in 600 ohms = zero level
−10	0.1000	0.31623	0.24495	20	100.00	10.0000	7.7461
− 9	0.1259	0.35481	0.27483	21	125.89	11.220	8.6912
− 8	0.1585	0.39811	0.30839	22	158.49	12.589	9.7514
− 7	0.1995	0.44668	0.34599	23	199.53	14.125	10.941
− 6	0.2512	0.50119	0.38820	24	251.19	15.849	12.276
− 5	0.3162	0.56234	0.43560	25	316.23	17.783	13.775
− 4	0.3981	0.63096	0.48875	26	398.11	19.953	15.459
− 3	0.5012	0.70795	0.54840	27	501.19	22.387	17.341
− 2	0.6310	0.79433	0.61527	28	630.96	25.119	19.457
− 1	0.7943	0.89125	0.69035	29	794.33	28.184	21.831
0	1.0000	1.00000	0.77461	30	1,000.00	31.623	24.495
+ 1	1.2589	1.1220	0.86912	31	1,258.9	35.481	27.484
+ 2	1.5849	1.2589	0.97514	32	1,584.9	39.811	30.837
+ 3	1.9953	1.4125	1.0941	33	1,995.3	44.668	34.600
+ 4	2.5119	1.5849	1.2276	34	2,511.9	50.119	38.822
+ 5	3.1623	1.7783	1.3775	35	3,162.3	56.234	43.560
+ 6	3.9811	1.9953	1.5459	36	3,981.1	63.096	48.875
+ 7	5.0119	2.2387	1.7341	37	5,011.9	70.795	54.840
+ 8	6.3096	2.5119	1.9457	38	6,309.6	79.433	61.527
+ 9	7.9433	2.8184	2.1831	39	7,943.3	89.125	69.035
+10	10.0000	3.1623	2.4495	40	10,000.00	100.000	77.461
+11	12.589	3.5481	2.7484	41	12,589.2	112.20	86.912
+12	15.849	3.9811	3.0837	42	15,848.9	125.89	96.698
+13	19.953	4.4668	3.4600	43	19,952.6	141.25	109.41
+14	25.119	5.0119	3.8822	44	25,118.9	158.49	122.76
+15	31.623	5.6234	4.3560	45	31,622.8	177.83	137.75
+16	39.811	6.3096	4.8875	46	39,810.7	199.53	154.59
+17	50.119	7.0795	5.4840	47	50,118.7	223.87	173.41
+18	63.096	7.9433	6.1527	48	63,095.7	251.19	194.57
+19	79.433	8.9125	6.9035	49	79,432.7	281.84	218.31

The VU meter is available commercially from several manufacturers and because of its moderate speed of action it is quite readable. It is considered today as the standard monitoring instrument and is available in several sizes, the one in most common use being approximately 4 in. square and available either with or without internal illumination.

The VU meter is also finding favor in applications other than across a-f lines, for example, as a monitoring instrument in sound recording on film, tape, or disk. It has largely superseded the previously used db meters having other dynamic characteristics, zero levels, etc. There is still some application, however, for the high-speed variety of db meter in the cutting of records where instantaneous indication and control of top level are desirable to prevent overcutting.

Table 9 is a useful tabulation of power levels, ratios, and voltages, all in terms of a zero level of 1 mw, with the last column—voltage—based on 1 mw in 600 ohms as the zero level. When interpreted on a VU meter as described, the values of db above and below this level also represent VU.

MEASUREMENTS OF PULSATING CURRENTS AND POTENTIALS

In making measurements of current and voltage which are neither true a.c. nor d.c., care must be taken to make the measurement with the correct type of instrument in order that a measurement be had of the actual value required.

37. Rectified-current Measurements. Rectified current, which may or may not be filtered, should in general be measured with a moving-coil permanent-magnet type of d-c instrument. This gives the average value. It is the value of current or voltage of interest when charging a battery and in general is the value of interest in vacuum-tube technique. Iron-vane and electrodynamometer instruments indicate the rms value which is used for determining the heating effect.

38. Measuring Pulsed Currents. Direct-current ammeters and milliammeters used to measure the average value of pulsed currents indicate the *average value* of the current but are heated by the *rms value*. For example, a 100-amp d-c pulse, 1 millisec long, repeated every 100 millisec, has an average value of 1 amp and will so indicate on a 1-amp d-c instrument. But, whereas a continuous current of 1 amp will cause a loss of 0.05 watt in a typical 1-amp instrument, the pulsed current in the example, giving the same reading, will cause a loss 100 times as great, or 5 watts; this is effectively the same loss as 10 amp of continuous current or a current overload of 10 times. This may overheat the instrument and cause permanent damage; in extreme cases instruments in applications of this kind have been burned out when reading below full scale.

Thus, for pulsed-current applications where the pulse repetition time is more than ten times the pulse length, special precautions should be taken. Oversize external shunts, heavy wire actuating coils, and other expedients usually allow an appropriate instrument design for any given set of conditions. The point must be stressed that standard instruments may be damaged in such service and the instrument manufacturer should be called on for an appropriate special design to meet the stated conditions.

To measure the a-c component of voltage, a capacitor may be placed in series with an a-c voltmeter of suitable range; the d-c component is blocked and the a-c value only is measured. The impedance of the capacitor at the frequency used (120 cycles for a full-wave rectifier system) should not be greater than 10 per cent of the instrument resistance; the impedances being in quadrature, the resulting error will be under 1 per cent. This is the simplest method of measuring hum in a rectified plate supply. Because of its high resistance, the rectifier voltmeter described previously is most satisfactory for this purpose.

Peak voltages and currents are best measured through the use of a vacuum-tube voltmeter with a large capacitance shunted by an extremely high-resistance d-c voltmeter. A cathode-ray oscillograph is also useful in such studies.

VOLTAGE-MEASURING INSTRUMENTS

39. Use of Current Meters to Indicate Voltage. All current-measuring instruments having a sensitivity in milliamperes may, with the addition of suitable series resistance, be used to indicate potential. The current drain of the instrument must be sufficiently low to abstract negligible energy from the circuit, as otherwise corrections must be made. With modern instruments of high sensitivity this requirement can usually be met.

Direct measurements of voltage are obtainable through electrostatic means, but the instruments are of limited utility because of their low torque and because the minimum ranges are rarely under several hundred volts. They are essentially instruments for the research laboratory.

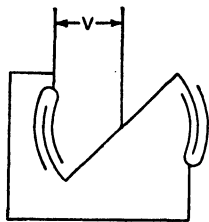



FIG. 22. Suspended-vane meter.

Electrostatic voltmeters depend on the attractive force which exists between two conducting plates between which a difference of potential exists. In their simplest form, the force of attraction between a stationary and a movable disk is balanced by a calibrated spring. The *Kelvin absolute electrometer* is constructed in this manner. The force of attraction is proportional to the square of the difference of potential between the plates. Such meters give the same indication on steady and alternating voltages and have neither wave-form nor frequency error.

One type of construction, used in suspended-vane meters, is shown in Fig. 22. The stationary plates are sections of two concentric cylinders, into which the cylindrical rotor turns. With the opposite poles of a magnet placed outside the stator plates, satisfactory damping is obtained from the currents induced in the loop. This type of construction is that used in the *Ayrton-Mather electrostatic voltmeter* built by the Cambridge Instrument Company.

Electrostatic voltmeters are very useful because of their high resistance and low power consumption at low frequencies. They cannot be used on high voltage at frequencies much above 1 Mc, because of the rapid increase of the power loss in the necessary insulation. This loss increases directly as the first power of the frequency and the square of the voltage. A hard-rubber insulator with a power factor of 0.004 and capacitance of 10 $\mu\mu$f will have, at a frequency of 10 Mc and voltage of 2.5 kv, a charging current of 1.5 amp and a power loss of 15 watts, both of which values are excessive.

MEASUREMENT OF RESISTANCE, IMPEDANCE, FREQUENCY

40. Resistance. While bridge measurements of resistance give greatest accuracy, direct-reading instruments are much used because there is no requirement for the manipulation of the controls, and they are widely used in production testing of resistance units as well as in general laboratory practice where the highest accuracy is not essential.

41. Ohmmeter. The simplest direct-reading ohmmeter consists of a milliammeter, a dry-cell battery, and a series resistance as shown in Fig. 23. With the terminals short-circuited, *i.e.*, $R = 0$, the shunt rheostat B is adjusted until the milliammeter I indicates full scale. When a resistance is connected across the terminals the reading of the instrument in terms of full-scale current can be expressed as

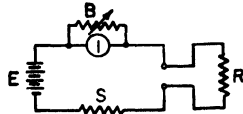



FIG. 23. Direct-reading ohmmeter circuit.

$$I \text{ (per cent)} = 100S/(S + R)$$

In this equation S is the resistance of the entire metering circuit to the terminals including both the series resistance marked S, any minor resistance in the battery proper, and the resistance of the shunted meter. Adjusting the meter shunt to take

care of variations in battery potential is preferred to the use of a variable series resistance since the shunt adjustment affects the circuit resistance to a much lesser degree. Further, to cover a battery-voltage variation of from 1.55 down to 1 volt per cell usually requires that the resistance B vary from 1 to 3.5 times the meter resistance; still better coverage is by making the B only the variable part or 2.5 times the meter resistance in series with a fixed resistance equal to the meter resistance. Thus, for a typical 1-ma meter having a resistance of 100 ohms the shunt would consist of a fixed resistance of 100 ohms and a variable resistance of 250 ohms.

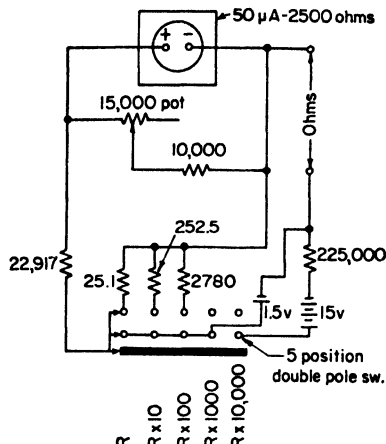



Fig. 24. Circuit for five-range series ohmmeter, with basic scale 0 to 500 ohms, 25 ohms at center.

With this type of ohmmeter the center-scale value is equal to the resistance of the shunted meter plus the series resistance. The total usable range may be taken as from 0 to 20 times the meter resistance. In this example, for measuring low resistance

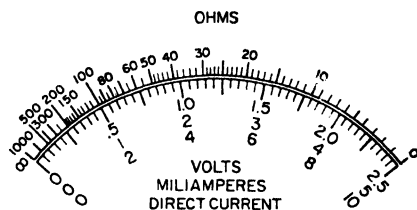



Fig. 25. Scale for ohmmeter of Fig. 24 with supplementary lower scale in volts and milliamperes.

values with a single dry cell, the shunted-meter resistance is approximately 53 ohms, the series resistance should then be 747 ohms, and the equivalent network resistance and the value for center scale would be 800 ohms.

The circuit of Fig. 24 shows a typical five-range ohmmeter obtained by switching in different shunt and series resistors, with a 15-volt battery of small current capacity for the highest range. The fixed shunt resistor is 4 times the meter resistances, with the variable unit 6 times the meter resistance. This means a minimum battery voltage of 1.375 volts per cell, but this higher cutoff point allows better use of the available meter sensitivity. The resulting scale for the lowest range of the diagram is shown in the upper scale of Fig. 25; multipliers are applied to this ohm scale for the other ranges as indicated. The evenly divided scales below the ohm scale are for voltage ranges of a typical volt-ohmmeter, obtained by switching from the ohmmeter network to an appropriate network of series resistors.

Fig. 26. Ohmmeter of Evershed and Vignole.

The upper limit of resistance measurements by this means depends upon the instrument sensitivity and battery voltage; a 50-μa instrument at 15 volts gives an excellent deflection on several megohms. The lower limit, since a minimum battery voltage of 1.5 volts must be used, is dependent only on the current capacity of the battery and the resistance of the leads. In general, for accurate work, the effective battery resistance must be calculated into the circuit as a part of the total series resistance.

The readings of an ohmmeter may be made independent of the applied voltage by dispensing with the controlling springs and obtaining the controlling torque from

a separate coil connected across the supply voltage. Figure 26 shows the circuit used by Evershed and Vignole in their ohmmeters of this type.

This construction was first used by Evershed for an ohmmeter designed to measure resistances up to 100 megohms. The source of voltage was a self-contained high-voltage magneto generator, giving voltages up to 500 volts. It was called a *Megger*. The same principle has now been applied to ohmmeters of lower range using battery voltages. The resistance range extends from 1 ohm to 5,000 megohms.

The upper resistance limit is greatly increased by replacing the microammeter with a vacuum-tube voltmeter connected across the fixed resistance, which may then have a maximum value of 10^9 ohms. The range will be 20,000 megohms.

The range is still further increased by using, instead of a VTVM, a degenerative d-c amplifier with a sensitivity of 0.1 volt for full-scale deflection of the indicating milliammeter. For a fixed resistance of 1,000 megohms, the current corresponding to this full-scale deflection is 100 $\mu\mu$a and 1 per cent of this, or 1 $\mu\mu$a, can be detected. At 100 volts the corresponding resistance values are 1 and 100×10^{12} ohms, since at all times the fixed resistance is negligible compared to the unknown resistance. It follows then that the voltage across the unknown resistor is essentially constant. The power supply is usually tube-regulated and is provided with a protecting resistor to limit the current on a short circuit to a safe value. One decade of resistance covers 90 per cent of the scale length, so that it is necessary to provide at least two fixed resistances per decade in order to keep the fractional accuracy of reading reasonably constant.

42. Measurement of Impedance. When the voltmeter-ammeter method is used with a source of alternating voltage, the ratio of voltage to current gives the impedance of the load

$$Z = E/I \qquad (2)$$

With the usual a-c instruments the corrections for the instruments are larger than for d-c instruments and more difficult to make because of their reactance. The high-resistance rectifier voltmeter and vacuum-tube voltmeter eliminate this difficulty.

The separation of impedance into its components requires the use of a wattmeter. The connections of Fig. 27*a* are usually used when no correction for instrument errors is to be made, while those of Fig. 27*b* allow the correction to be made quite easily. For this distinction the current coil of the wattmeter is grouped with the ammeter and its potential coil with the voltmeter. As before, the impedance of the load is given by Eq. (2). Its power factor is the ratio of the wattmeter readings to the product of voltage and current.

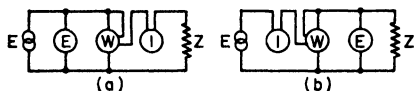



FIG. 27. Measurement of impedance.

$$\text{Pf} = \cos\theta = W/EI \qquad (3)$$

where θ is the phase angle between voltage and current. The resistance of the load is

$$R = W/I^2 \qquad (4)$$

and the reactance

$$X = \sqrt{W^2 - R^2} \qquad (5)$$

With the knowledge as to whether the load is inductive or capacitive, its inductance or capacitance may be calculated from

$$X = \omega L = -1/\omega C \qquad (6)$$

where $\omega = 2\pi f$.

43. Measurement of Capacitance. Since the power factor of the usual capacitor is small, its reactance is approximately equal to its impedance. This may be measured directly by the voltmeter-ammeter method and the capacitance calculated from Eq.

(6). At a given voltage and frequency, a single ammeter reading is sufficient, and the ammeter may be calibrated to read capacitance directly.

Capacitance may also be measured on a single indicating meter whose readings are independent of the applied voltage. The moving element consists of two coils set at

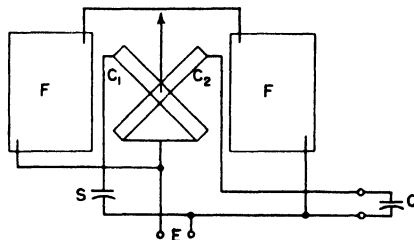



FIG. 28. High-frequency microfarad meter.

right angles to each other. There are no controlling springs. The connections used in the h-f Weston microfarad meter are shown in Fig. 28.

References

Brown, H. A.: "Radio-frequency Electrical Measurements," McGraw-Hill, 1938.
Cambridge Instrument Company, Catalogue.
Drysdale and Jolley: "Electrical Measuring Instruments," Part I, Van Nostrand, 1924; also 2d ed., revised by G. F. Tagg, Wiley, 1952.
General Radio Company, Catalogue.
Hague, B.: "Alternating-current Bridge Methods," Pitman, London, 1938.
Harris, F. K.: "Electrical Measurements," Wiley, 1952.
Hund, A.: "High-frequency Measurements," McGraw-Hill, 1933.
Kennelly, A. E.: "Electrical Vibration Instruments," Macmillan, 1923.
Laws, F. A.: "Electrical Measurements," McGraw-Hill, 1938.
Leeds & Northrup Company, Catalogue.
Miller, J. H.: "Instruments, Electrical Measuring," Encyclopaedia Britannica, 1956.
Moullin, E. B.: "Radio Frequency Measurements," Griffin, 1926.
Terman, F. E.: "Measurements in Radio Engineering," McGraw-Hill, 1935.
Weston Electrical Instrument Corporation, Catalogue.

CHAPTER 14

ALTERNATING-CURRENT MEASUREMENTS

BY B. J. DASHER[1] AND JOHN TAYLOR[1]

VACUUM-TUBE VOLTMETERS

At frequencies above a few hundred cycles dynamometer and iron-vane ammeters become impractical, partly because at higher frequencies greater sensitivity is required (because impedances are likely to be higher) and partly because stray capacitance and other second-order effects make it difficult to maintain reasonable accuracy. At these higher frequencies thermocouples offer the only practical means for measuring current. They are satisfactory up to frequencies of the order of 100 Mc. Above this range indirect methods must be used. For example, current can be computed from measurements of voltage and impedance.

Fortunately, it is comparatively easy to measure voltage at frequencies up to several hundred megacycles. Vacuum-tube voltmeters are most often used for this application. Rectifiers in conjunction with sensitive milliammeters are also frequently used.

1. D-c Vacuum-tube Voltmeters (VTVM). A simple form of vacuum-tube voltmeter consists of a triode amplifier with a milliammeter in the plate circuit as shown in Fig. 1. If the supply voltages are selected so that the operating point for the tube is in the "linear region" the average plate current is very nearly proportional to the input voltage E_{in}. Since the plate current is not zero when E_{in} is zero, it is desirable to provide a "zero balance." A convenient way to do this is shown in Fig. 2.

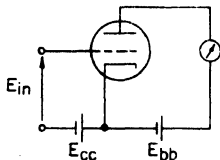



FIG. 1. D-c vacuum-tube voltmeter.

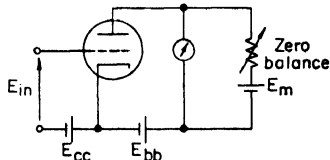



FIG. 2. Vacuum-tube voltmeter with zero balance.

The battery E_m tends to send current through the meter in a direction opposite to that of the plate current. When the current in the meter is zero, the voltage drop in the balancing resistor just equals E_m. The sensitivity of the device is reduced by the addition of the zero balance because the incremental plate current must divide between the meter and the zero-balance circuit. However, the sensitivity of the meter can be increased to compensate for the reduction in inherent sensitivity.

The maximum sensitivity that can be obtained with this basic circuit is limited by the stability that can be achieved in balancing out the zero-signal plate current. Fluctuations caused by variations of heater voltage, plate-supply voltage, contact potentials, etc., must not be permitted to produce meter deflections comparable in magnitude to the deflections produced by the voltage to be measured.

Another basic VTVM circuit is shown in Fig. 3. In this arrangement, a voltmeter is used as the cathode resistor in a cathode follower.[1*] As in the previous circuit, a zero balance can be added to compensate for the zero-signal plate current. One such scheme is shown in Fig. 4. The input impedance of this circuit is extremely high—hundreds or thousands of megohms. Also, the degenerative feature inherently tends

[1] Georgia Institute of Technology.
* Superior numbers refer to references at the end of the chapter.

to stabilize the circuit. The range can be changed merely by changing the range of the voltmeter. In Fig. 3 the incremental current in the tube is given by

$$i_p = \frac{\mu E_{\text{in}}}{r_p + \mu R_m} = \frac{E_{\text{in}}}{R_m} \frac{1}{1 + r_p/\mu R_m} \tag{1}$$

in which μ is the amplification of the tube, r_p is the plate resistance of the tube, and R_m is the resistance of the voltmeter. Equation (1) may also be used for Fig. 4, provided R_m is replaced by the equivalent resistance of the meter and zero-balancing resistor is parallel.

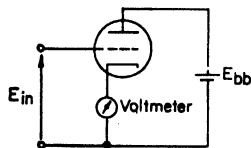



FIG. 3. Degenerative or cathode-follower circuit.

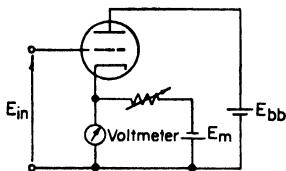



FIG. 4. Degenerative circuit with zero balance.

Variations of the circuit in Fig. 4 are widely used in d-c voltmeters and also as basic parts of a-c voltmeters. The latter are discussed in Sec. 5.

2. Diode Voltmeters. A diode voltmeter is a form of rectifier instrument. However, the term "rectifier" usually is applied to instruments employing copper-oxide rectifiers whereas the term "diode" usually implies the use of vacuum diodes as rectifiers. In addition, diode voltmeters are generally designed to give readings proportional to the peak value of an a-c voltage, but rectifier instruments give readings proportional to the full-wave average values. Because of the characteristics of copper-oxide rectifiers, rectifier instruments are restricted to use at audio frequencies, but diode voltmeters may be used at frequencies of the order of several hundred megacycles. Semiconductor diodes are sometimes used instead of vacuum diodes. The indicating instrument is invariably a D'Arsonval-type d-c meter.

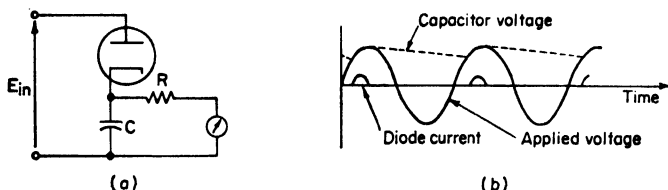



FIG. 5. (a) Basic diode circuit; (b) wave forms.

The basic circuit of a diode voltmeter is shown in Fig. 5a. The corresponding current and voltage wave forms are shown in Fig. 5b. If the time constant RC is large compared with the period of the a-c voltage to be measured, the voltage across the capacitor and, thus, across the resistance will be essentially a d-c voltage equal to the positive peak of the input voltage. Note that the diode is in series with the input voltage. Hence, the source must provide a d-c path; otherwise no current can flow through the diode. If the input voltage contains a d-c component, the meter current indicates the net peak positive voltage.

A second circuit, sometimes called a shunt circuit, is shown in Fig. 6a and the corresponding wave forms are shown in Fig. 6b. Compared with the series circuit, the shunt circuit has the advantage that the measured source need not provide a d-c path. Also, it is frequently advantageous to have the cathode of the diode at ground potential. The average voltage across the diode and, thus, across the meter resistance is nearly equal to the positive peak value of the input voltage, exclusive of any d-c component, since the capacitor C prevents any direct current from flowing in the meter circuit. The full input voltage is impressed on the meter-resistance circuit in addition to the d-c voltage.

In both the above circuits the emf induced by the emission velocity of the electrons from the cathode of the diode will produce a small deflection of the meter when the input voltage is zero. This emf is normally about 1 volt. This emf may be balanced out by means of a zero-balance circuit similar to those already described. Sometimes a mechanical adjustment is provided on the meter movement for zero balance.

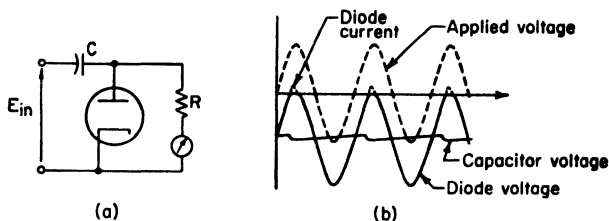



FIG. 6. (*a*) Shunt circuit; (*b*) wave forms.

For small voltages (10 volts and less) the response of the diode voltmeter is noticeably nonlinear and special scales must be provided for the low-voltage range.

As may be seen from Fig. 6*b*, the peak voltage across the resistor is the peak-to-peak voltage of the input wave. Thus, if the resistance-meter combination is replaced by the peak reading meter (Fig. 5*a*), the so-called *voltage-doubling* circuit of Fig. 7 is obtained. In this circuit the first diode has been reversed so as to give a positive direction for the d-c current. In these discussions the voltage drop in the diode during conduction has been neglected. Although this voltage drop causes the voltage across the resistance to be slightly less than otherwise, the current is almost directly proportional to the input voltage. Thus, the meter can be calibrated to indicate the actual voltage.

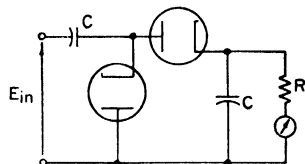



FIG. 7. Voltage-doubling circuit.

3. Input Impedance of Diode Voltmeter. It is always wise to be aware of the finite input impedance of voltmeters, especially at high frequencies. The presence of the meter may appreciably affect the circuit under measurement. The input impedance is represented with good accuracy as a capacitance in parallel with a resistance. The capacitance can be reduced to 1 or 2 $\mu\mu$f with careful design, but it is usually much larger. When the voltage being measured is sinusoidal or nearly so, the input resistance of the instrument can be evaluated in terms of the power dissipated in the resistor. In the circuit of Fig. 5,

$$R_{\text{in}} \approx R/2 \tag{2}$$

In Fig. 6,

$$R_{\text{in}} \approx R/3 \tag{3}$$

and in Fig. 7,

$$R_{\text{in}} \approx R/8 \tag{4}$$

The fundamental component of the diode current leads the input voltage slightly. The angle is small and decreases with increasing frequency. When the voltage is not sinusoidal the power input is approximately $(E_{\text{peak}})^2/R$ and this is usually a more significant quantity than the equivalent input resistance since the value of the resistance depends on how it is defined. At frequencies high enough so that the transit time for the electrons in the diode is an appreciable part of a period, the input resistance is less than given by the appropriate equation (2), (3), or (4) because of the energy absorbed by the electrons. A careful analysis of this effect can be made only by considering specific circuits, tubes, etc. Transit-time effects also introduce additional errors.[2]

4. Wave-form Considerations. Vacuum-tube voltmeters are usually calibrated to indicate the rms value of a sinusoidal voltage. If the wave form is not sinusoidal the

indications must be interpreted with due attention to the wave shape. It is essential to know the circuit employed in the voltmeter. In the case of a square wave, the peak value is found by multiplying the meter reading by $\sqrt{2}$. For unsymmetrical waves useful generalizations are difficult to make. For example, in the case of very short pulses, the current pulse through the diode may be quite large, thus causing an excessive voltage drop in the diode. Under these circumstances, the meter reading will be appreciably less than the voltage peak. A full-wave circuit, such as Fig. 7, will give the correct value of the fundamental component of a wave containing only even harmonics as shown in Fig. 8. Such wave forms are typical of those produced by distortion in vacuum-tube amplifiers. On the other hand, if only odd harmonics are present, the wave form may be peaked as in Fig. 9. In this case, the reading will be appreciably higher than the fundamental component.

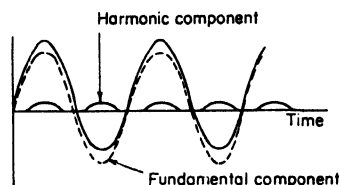



FIG. 8. Voltage wave containing only even harmonics.

These "errors" caused by nonsinusoidal wave forms are not so serious as they may at first appear. Thus, in the case of pulses, the peak amplitude may be the desired quantity. Similarly, it is usually satisfactory to measure the peak value of a saw tooth, square wave, or other wave of known shape. But if the wave-form characteristics are not known, the voltmeter readings may be quite meaningless. A cathode-ray oscilloscope is indispensable when dealing with nonsinusoidal waves. Its use as a voltmeter is discussed in Sec. 9.

FIG. 9. Voltage wave containing only odd harmonics.

5. Rectifier-amplifier Voltmeter. The sensitivity of the diode voltmeters described in Sec. 2 is limited by two factors: (1) for voltages below 10 volts the scales are nonlinear and dependent on individual tube characteristics; (2) full-scale current must be at least 50 μa unless fragile meter movements are used. Thus, the maximum practical sensitivity is of the order of 10,000 ohms per volt. To permit higher input impedances to be obtained and to permit very small voltages to be measured rectifier-amplifier combinations are commonly employed in commercial instruments. Two basic arrangements are in general use. In one, a diode rectifier is used followed by a d-c voltmeter. In this case, the voltage developed by the rectifier is used as the input to the d-c meter. Since no meter is needed in the diode circuit, the resistance can be made 10 to 100 times larger than would otherwise be possible.

In the second type of rectifier-amplifier instrument, the amplifier is an a-c amplifier and a diode voltmeter is connected to its output.

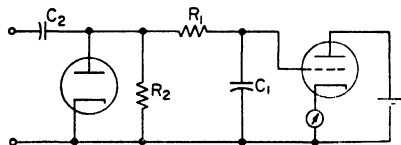



FIG. 10. Diode rectifier and d-c voltmeter.

Figure 10 shows a simplified schematic diagram of an instrument of the first type. R_1 and C_1 constitute a filter to prevent the a-c component of the diode voltage from appearing on the grid of the triode. In Fig. 11 a second diode-triode combination is added to provide a balance circuit which has much greater stability, since variations caused by supply-voltage variations, aging of tubes, and the like tend to cancel. The sensitivity of these circuits can be increased by the addition of a d-c amplifier stage but this does not overcome the nonlinear characteristics of the diode at low input voltages.

In the second type of circuit, the amplifiers are a-c amplifiers. In addition to being less critical as regards drift and aging of tubes, these circuits provide a comparatively

large voltage for the diode circuit. Various modifications of the circuits previously discussed may be used as the rectifier. In Fig. 12 the full-wave circuit shown yields a reading proportional to the full-wave average. Resistor R_2 prevents the capacitor from charging to the peak voltage.

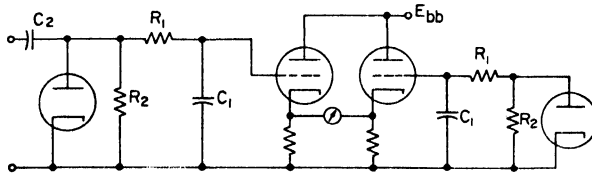



FIG. 11. Balanced voltmeter circuit.

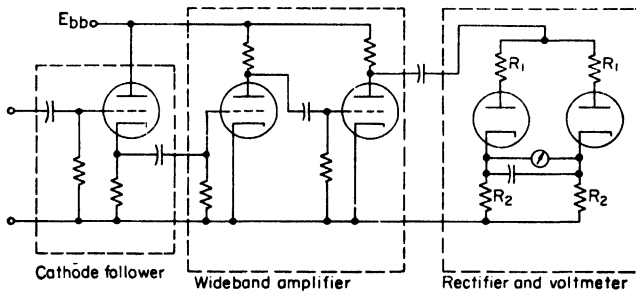



FIG. 12. Amplifier-rectifier vacuum-tube voltmeter.

The upper frequency limit of this circuit depends primarily on the band width of the amplifier. Practical designs give calibrations substantially independent of frequency up to 10 Mc. At higher frequencies, it is desirable to install the cathode follower in a probe so that the leads associated with the high-impedance portions of the circuit may be as short as possible. For the same reason, the diodes used in instruments of the type in Fig. 10 are frequently installed in a probe. Since these instruments do not require wide-band amplifiers, they are, in general, better suited for very high frequencies, but as pointed out above, they do not yield as high sensitivity as the second type. For frequencies above 100 Mc, the vacuum diodes may be replaced by semiconductor crystal diodes.

6. Square-law Voltmeters. Many variations of the circuits already described are possible,[2] but these circuits include the basic principles of most commercial instruments presently available. One other circuit is sometimes useful in the laboratory, although it is not generally available in a commercial instrument. This is the square-law voltmeter. The plate current of a triode operating near cutoff is nearly proportional to the square of the grid voltage. Hence, when an a-c voltage is applied to the grid, the average current, which is the d-c plate current, is proportional to the square of the rms grid voltage. Figure 13 shows a circuit that can be used as a square-law voltmeter. For large input voltages, this circuit operates essentially as a plate-circuit rectifier and gives approximately linear rectification and an indication proportional to the half-wave average of the input voltage.

FIG. 13. Square-law voltmeter with balancing tube.

7. Slide-back Voltmeters. The slide-back voltmeter is useful for measuring d-c voltages. It can be designed for greater sensitivity, higher input resistance, and

greater accuracy than other types of VTVM, but it has the disadvantage of being basically a comparison device and, consequently, not direct-reading.

The principle is illustrated in Fig. 14. This circuit is in reality an electronic potentiometer.[3] The resistor R_s is a calibrated "slide-wire" potentiometer, which, for best accuracy, should be of the Thomson-Varley type. If the total resistance of R_s is, say, 100,000 ohms then a current of 1 ma gives a total voltage of 100 volts across R_s. To operate the instrument, the push button Pb is depressed and R adjusted so that the tube current is exactly 1 ma. Then the push button is released and the potentiometer adjusted so that the current is again exactly 1 ma. Now the drop E_s across the potentiometer is exactly equal to the unknown voltage. By carefully selecting the operating point of the tube, the grid current may be reduced to a few parts in 10^9 or less to yield an extremely high input impedance. Voltages can be measured to four significant figures without difficulty. Of course, the slide wire must have corresponding precision, but the meter need be accurately calibrated at only one point.

FIG. 14. Slide-back vacuum-tube voltmeter.

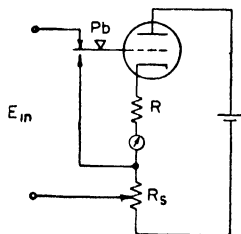
When less precision is required, E_s may be measured with a voltmeter instead of a potentiometer.

8. Accuracy. The accuracy of vacuum-tube voltmeters is less than that of good-quality instruments for use at low frequencies. The reason for this is too involved to discuss here. However, accuracy attainable with commercial instruments is usually as good as 5 per cent and sometimes better. It is seldom better than 1 per cent, but accuracy better than this is seldom required. Frequency errors are negligible over a wide range. Therefore, a tube voltmeter can be calibrated at 60 cps by means of a standard instrument. In this way, higher accuracy can be obtained.

9. Cathode-ray Oscilloscope. A cathode-ray oscilloscope serves as a convenient voltmeter, especially at frequencies in the megacycle range. Some instruments have calibrated attenuators and self-contained calibrating circuits. Otherwise, the oscilloscope can be used as a comparison device, using a low-frequency instrument as a reference. It is necessary, of course, to have a knowledge of the frequency-response characteristics of the oscilloscope to obtain reasonable accuracy in this way.

10. Ground and Capacitance Considerations. When working at high frequencies it is not enough to have a well-constructed instrument. Failure to observe a few simple precautions can lead to errors more serious than those caused by the limitations of the meter.

One very important factor is the matter of ground connections. The ground terminal of the VTVM must always be connected to the ground terminal of the circuit under test. Otherwise, the capacitance between the two ground circuits will be added to the test circuit between its ground terminal and the point to which the meter ground is connected. Unless this part of the circuit has a very low impedance, spurious readings will result. This sometimes renders it impossible to measure the desired voltage. For example, the voltage across an interstage coupling capacitor usually cannot be measured directly. Instead, each voltage must be measured with respect to ground and the difference taken, with due attention to the phase difference in the case of alternating voltages.

Another factor that must always be remembered, especially in r-f circuits, is the influence of the meter on the circuit under test. At best, the VTVM adds a few micro-microfarads across the test points. This additional capacitance may alter the voltage. Resonant circuits, for example, should usually be retuned after the voltmeter is connected. The added capacitance may cause high-gain amplifiers to oscillate or it may otherwise alter their performance. As a general rule, the leads between the voltmeter and the test circuit should be as short as possible and when a probe is used the shielded cable should be kept away from high-potential portions of the test circuit. Excessively long leads may also produce resonance effects causing the voltage at the meter terminals to be different from the desired voltage.

FREQUENCY, WAVE FORM, AND PHASE

The fundamental frequency standard is the period of rotation of the earth. This period can be measured with great accuracy by astronomical techniques and serves as a reference to which all frequency measurements must ultimately be compared.

Standard frequency generators are classified as primary or secondary. A primary standard is one arranged for direct comparison with the earth's period at regular intervals. A secondary standard is one whose calibration is obtained by reference to a primary standard. The stability and accuracy of present frequency standards is excellent and it is relatively easy to measure frequency to an accuracy of a few parts in 10^8.

11. Frequency Standards. Present-day primary-standard frequency generators usually employ a carefully designed crystal-controlled oscillator to generate a frequency of 100 kc. This frequency is reduced to 1,000 cps by frequency dividers so designed that the 1,000-cycle frequency is used to drive an electric clock whose time is compared with the earth's period. Crystal oscillators are available that have a long-time frequency stability of a few parts in a hundred million. This represents approximately 0.001 sec in 1 day.

Secondary standards also employ carefully designed crystal-controlled oscillators. They differ from primary standards principally by the omission of the clock and associated means for comparing it with observatory time. Such a secondary standard can be expected to have a stability of a few parts in a million for long periods of time.

For laboratory standards, the National Bureau of Standards standard frequency transmissions are convenient. By means of a good communications receiver transmissions can be received in most parts of the United States 24 hr a day. They provide frequencies with an accuracy of better than 1 part in 50,000,000.

Time is also announced every minute from the Canadian station CHU in the 3- and 7-Mc bands.

Signals from commercial broadcast radio stations are usually maintained accurate to a few parts per million.

12. Frequency Meters. Two types of direct-reading frequency meters are presently available for use at radio frequencies.

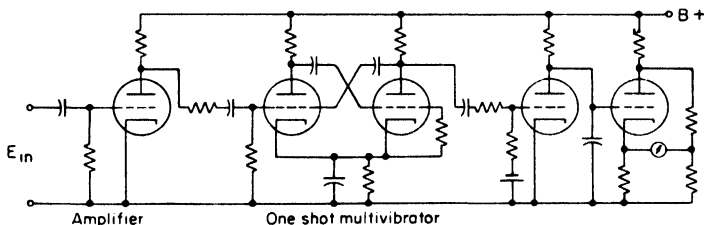



FIG. 15. Counting-rate meter or frequency meter.

One type using direct indicating instruments in conjunction with electronic circuits covers a useful range from 3 cps full-scale to 60 kc full-scale. The electronic circuit generates pulses which are all of identical shape but which depend entirely for their frequency on the frequency of the applied voltage. The identical pulses are rectified and "counted" on a d-c instrument whose deflection is proportional to the number of pulses per second. The instrument is calibrated in terms of frequency in cycles per second. The maximum frequency range that can be covered with this type of instrument is limited only by the rate at which the pulses can be generated. The accuracy is similar to that of other direct-reading electronic instruments—about 1 per cent.

Figure 15 shows a simplified schematic diagram of the *counting-rate* type of frequency meter.[4] The input signal is amplified and applied to the input of a modified Eccles-Jordan circuit ("one-shot" multivibrator), which generates a standard pulse for each positive half

of an a-c voltage or each positive pulse input. The wave shape of this pulse is independent of its rate, which is the same as the input pulse rate. The positive pulses from the pulse generator are applied to the grid of a tube that has been biased beyond cutoff; plate current flows only while the positive pulses are applied. A filter or averaging circuit delivers a d-c voltage to a VTVM. Since the amplitude and duration of the pulses are constant and their separation equals the period of the input frequency, the average output voltage is proportional to the input frequency.

A second type of direct-reading frequency meter is literally a counting instrument.[5] A block diagram of a typical circuit is shown in Fig. 16.

The signal whose frequency is to be measured is led first to a pulse-forming circuit that delivers a very short pulse each time the sign of the input signal voltage changes from negative to positive. The pulses are fed to a gated amplifier and then to an electronic counting circuit. The time-base generator consists of a very stable, crystal-controlled oscillator and suitable frequency dividers. Suppose the oscillator frequency is 100 kc. Then five 10 to 1 frequency dividers provide an accurate 1 cps output. When the "count" switch is depressed, the time-base generator turns the gated amplifier "on" at the beginning of the next 1-cps period and turns it "off" at the beginning of the next successive period. Thus, the counter is permitted to count pulses for exactly 1 sec and the number of counts displayed is the average frequency (averaged over 1 sec) of the input signal.

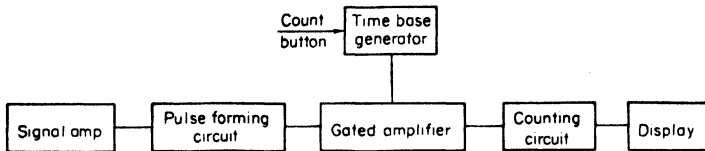



Fig. 16. Block diagram of counting-type frequency meter.

The accuracy of the pulse-counting-type frequency meter is limited by the stability of the timing oscillator and by the fact that an integral number of counts must be registered. There is an inherent uncertainty of 1 count in the frequency indicated. Thus, if a frequency of 100 cps is to be measured to an accuracy of 1 part in 10^6 it is necessary to count for 1,000 sec. For the same accuracy the same number of counts will be required regardless of frequency, but for higher frequencies the counting time is, of course, less.

In commercial instruments the timing oscillator usually has an accuracy of a few parts per million, but provision can be made to use a separate oscillator of higher precision. The frequency range extends to 1 Mc and can be made higher through the use of a heterodyne oscillator (see Sec. 13).

Frequency Comparisons. A very common way to measure frequency is to compare the unknown frequency with a standard frequency. The heterodyne frequency meter operates on this principle. The basic components of a heterodyne frequency meter are a stable, tunable, calibrated oscillator, a detector and means to observe the output of the detector. Signal amplifiers and other elaborations may be provided according to the particular intended use of the instrument but the principle of operation remains the same. The unknown signal and oscillator output signal are fed into the detector (sometimes called the mixer). The frequency of the output from the detector equals the difference between the known and the unknown frequency. When the oscillator is adjusted so that this difference is zero, then its frequency equals that of the unknown frequency. The detector output may be observed by means of headphones, loudspeaker, cathode-ray oscilloscope, or even a voltmeter. When the detector output frequency has been adjusted to "zero beat," the frequency of the unknown signal is read from the calibrated oscillator.

The accuracy of the heterodyne frequency meter is limited by the accuracy and stability of the local oscillator. It is very difficult to construct a precision oscillator whose frequency can be varied over a wide range. Hence, the accuracy of heterodyne

meters is seldom better than a few parts in 10,000. In addition to the limitation set by the stability of the oscillator, there is the problem of providing a calibrated scale that can be read to more than three or four significant figures. The oscillator frequency itself may drift with aging of components and it may vary appreciably as a result of variations of power-supply voltages, ambient temperatures, and other such parameters. Consequently, heterodyne frequency meters are commonly provided with reference oscillators, in the form of carefully designed crystal oscillators. These can be used to check the calibration of the instrument so that adjustments can be made to compensate for drift. When such a crystal calibrator is included the calibration can be made dependable to about 1 part in 10,000 for indefinitely long periods of time.

13. Interpolation Measurements. To make frequency measurements to an accuracy greater than 1 part in 10^6 it is usually necessary to employ a frequency standard and some form of interpolation procedure. The basic principle of interpolation measurements is to use a variable oscillator to cover the frequency range between two precise frequencies generated by a frequency standard. The precision required of the variable oscillator is then greatly reduced because it is only required to compare the unknown frequency with the difference between the two known frequencies.

To illustrate, let f_1 and f_2 be two standard frequencies separated by 1,000 cps and let f_x be an unknown frequency between f_1 and f_2. The variable oscillator may be adjusted for zero beat with each of these frequencies. Now the difference between the readings corresponding to f_1 and f_2 is known to be exactly 1,000 cps; so f_x can be calculated by proportion. If the stability of the variable oscillator permits the difference to be measured to 1 cps the over-all accuracy of the measurement is approximately 1 part in f_1 parts provided the accuracy of f_1 is at least as great.

In practice, many variations of this idea are used. A common procedure is to heterodyne (beat) the unknown frequency with the standard frequency and measure the difference with a heterodyne frequency meter. This is an effective way to extend the range of counting-type frequency meters as suggested in Sec. 12. Standard frequencies may be obtained from harmonics of the precision oscillator.

The accuracy of measurements involving heterodyne processes can be determined by computing the total uncertainty in cycles per second. As an example, suppose the reference frequency is 1 Mc, accurate to 1 part in 10^7. This corresponds to 1 cycle in 10 sec. Let the unknown frequency be approximately 1.05 Mc. The difference also has an uncertainty of 1 cycle in 10 sec. Suppose a 50-kc signal is available from the standard, obtained from the 1-Mc oscillator by frequency division. Its accuracy is 1 part in 10^7, or 0.05 cycle in 10 sec. Now suppose that heterodyning the first difference with the 50-kc standard gives a frequency of less than 1,000 cps which is measured with a heterodyne frequency meter as 570.1 cps. The heterodyne meter has an accuracy of 1 part in 10^4, which also amounts to 1 cycle in 10 sec. Hence, the total uncertainty is 2.05 cps in 10 sec and the measured frequency is 1,050,570.1 ± 0.205 cps.

The distinction between the stability and the accuracy required of the interpolation oscillator should be noted. The stability, *i.e.*, the ability to maintain a given frequency within narrow limits over long periods of time, must be such that the frequency does not change more than the permissible error during measurements. The accuracy, *i.e.*, the precision with which the frequency can be determined from the dial settings, need not be as great since the standard signals provide, in effect, an accurate calibration.

It is common practice for a primary or a secondary standard to provide known frequencies over a wide range for use as "markers." In one arrangement the master oscillator is crystal-controlled at 100 kc. Ten-to-one frequency dividers provide additional signals at 10 kc, 1 kc, and 100 cps. Each of these signals, in turn, is used to drive a harmonic generator, such as a multivibrator, and all these frequencies have the same accuracy as the master oscillator. Similarly, frequency multipliers can be used to generate accurate multiples of the 100-kc frequency.

At lower frequencies (below 1 Mc) multivibrators are very satisfactory as harmonic generators. At higher frequencies, crystal rectifiers (germanium or silicon) are generally more satisfactory. With careful design, harmonics including frequencies as high as 25,000 Mc can be generated.

Although a 1-kc multivibrator may provide harmonics at several megacycles, they are likely to be very weak. Also, in making measurements, it may be difficult to determine which harmonic is being observed. A more satisfactory way to obtain signals of 1-kc separation in the megacycle range is to modulate, for example, a 1-Mc signal by the 1-kc multivibrator signal. Then the side bands of the modulated carrier, produced by harmonics of the multivibrator signal, will be spaced exactly 1 kc apart.

One method of using such equipment is as follows: Fig. 17 indicates the relative positions of the known frequencies. A heterodyne frequency meter or a good communications-type receiver (provided with beat-frequency oscillator) is first tuned to the unknown frequency to locate it approximately. Then, with the modulation frequency f_1 turned off, the receiver is tuned to f_0. Next, the modulation is turned on and the receiver tuned back to f_x. The side bands can be "counted" as they are passed. In this way it is relatively easy to determine which side band lies closest to the unknown frequency. If the standard and the unknown signals are now impressed on a detector, the difference frequency will be obtained and this can be measured by any of the several means already described.

FIG. 17. Standard-frequency marker signals.

14. Comparison by Means of Cathode-ray Oscilloscopes. An oscilloscope affords a convenient means for comparing two frequencies, especially when their ratio is not large. Applying one signal to the horizontal plates and the other to the vertical plates produces a pattern known as a *Lissajous figure*. When the two frequencies are exactly equal, the pattern is a circle or ellipse that is perfectly stationary. Patterns corresponding to frequency ratios of 1:1, 2:1, and 3:1 are shown in Fig. 18. Provided the ratio of the two frequencies is that of two small whole numbers, the number of horizontal and the number of vertical excursions can be counted and the ratio then determined. When the ratio is large or involves large numbers, such as 17:21, it is not easy to count the excursions because of the rather complicated figures produced. It is nearly impossible to make the figures really stationary, and, consequently, only rather simple patterns are practical for most purposes. One of the frequencies must be variable to permit the ratio to be adjusted so that the pattern is at least nearly stationary. Also, relatively large voltages must be available because most oscilloscopes have limited amplification and this is sometimes a disadvantage. The method finds its greatest usefulness at audio frequencies, but with care it can also be used at much higher frequencies.

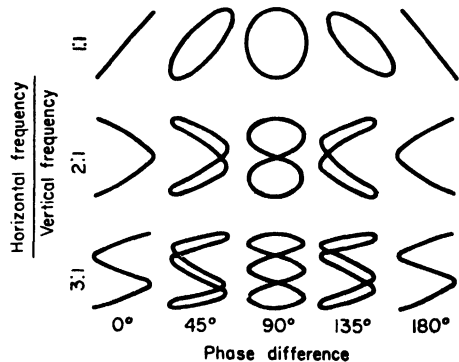



FIG. 18. Lissajous figures.

15. The Method of Sliding Harmonics. This is a novel use of a stable, tunable oscillator.[6] The principle is illustrated in the block diagram of Fig. 19. Multivibrators can be used as frequency dividers or multipliers. They will remain synchronized over the 1 per cent frequency variation illustrated in Fig. 19. The 100th harmonic of

1.01 Mc is the same frequency as the 101st harmonic of 1.00 Mc. The 101st harmonic of 1.01 Mc is slightly higher than the 102d harmonic of 1.00 Mc. Further examination shows that the 1 per cent variation permits continuous coverage of frequencies above the 100th harmonic, in this case above 100 Mc. Similarly, if a 100-kc multivibrator is employed, it will provide continuous coverage for frequencies above 10 Mc, etc. Thus, this system provides, in effect, continuous interpolation with an accuracy approaching that of the crystal-controlled oscillator.

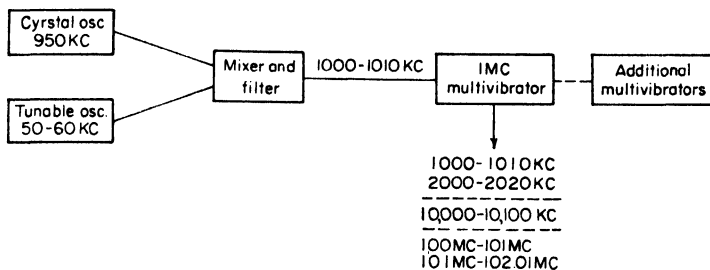



FIG. 19. Interpolating frequency standard using sliding harmonics.

16. Wavemeters. A high-Q tuned circuit provided with a resonance indicator makes a simple inexpensive frequency meter where great accuracy is not required. Such a device is commonly called a wavemeter. One form is represented by the circuit diagram in Fig. 20. The capacitance voltage divider permits the rectifier-microammeter to be loosely coupled to the resonant circuit so that the Q of the circuit will not be appreciably decreased. The variable capacitor is calibrated in terms of resonant frequency, and plug-in coils permit a wide range of frequencies to be covered with a single capacitor.

Wavemeters can be operated as either absorption or reaction devices. In the former case, the coil of the wavemeter is loosely coupled to the circuit whose frequency is to be measured and the capacitor is adjusted for maximum microammeter current. Since the resonant frequency of the wavemeter circuit is changed slightly by the presence of the circuits being measured, the coupling should be as loose as possible while still maintaining a satisfactory indication of resonance. (The resonant frequency of an oscillator circuit may also be affected by the presence of the wavemeter if they are too closely coupled.)

FIG. 20. Tuned-circuit-type wavemeter.

Grid-dip Meters. When used as a reaction device, the wavemeter is also loosely coupled to the circuit under test, but in this case resonance is determined by the reaction of the wavemeter on the test circuit. For example, the grid current of an oscillator will "dip" slightly as the wavemeter is tuned through the operating frequency; the voltage across an amplifier tank circuit will likewise decrease slightly. When used in this way, the wavemeter need not have its own resonance indicator.

The accuracy with which frequency can be determined by means of a wavemeter depends primarily on the Q of the wavemeter circuit and precision with which the capacitor dial can be calibrated. A carefully made worm-screw drive will permit readings to about 1 part in 10^3. Coil Qs from 100 to 500 can be obtained without great difficulty, the higher Qs usually being easier to obtain at higher frequencies. The "exact" maximum deflection is difficult to find because of the flat-top characteristics of resonance curves. For increased accuracy, reference readings should be taken at points off resonance where the characteristic has steep sides. For example, the settings corresponding to equal readings either side of resonance may be noted and the average of these two frequencies will give the desired frequency. (The geometric

mean of these two frequencies is theoretically a closer approximation, but the error introduced by taking the average is usually less than other uncertainties.) The half-current frequencies or the half-power frequencies (current approximately 70 per cent of maximum) make convenient reference points. Using these refinements the wavemeter will give an accuracy of about 0.1 per cent. It is difficult to achieve greater accuracy and an accuracy closer to 1.0 per cent is more usual. Of course, other factors, such as the stability of the tuned circuit with respect to temperature, also affect the accuracy of wavemeters but these factors are almost always of less importance than the limitations set by the accuracy of adjusting and reading the instrument.

In view of the much greater accuracy that can be obtained by other means, the wavemeter is not considered to be a precision instrument. Also, it is necessary to have enough power available to operate the resonance indicator. Greater sensitivity can be obtained by substituting a VTVM (which may be connected across the calibrated capacitor) for the rectifier-microammeter. However, wavemeters are most useful when working with oscillators or transmitters for rough measurements of frequency. Their ease of adjustment and freedom from ambiguities make them useful adjuncts to more precise equipment.

17. Standing-wave Techniques. The "lumped-parameter" wavemeters described in the previous section can be used at frequencies ranging from a few kilocycles to a few hundred megacycles. In the range from approximately 100 to 10,000 Mc standing-wave techniques are useful. Resonant transmission lines, called *Lecher wires,* are useful at the lower end of this range while coaxial lines are better suited to the higher end.

Resonant lines may be used in two ways. One of these is to connect a voltage or current indicator at some convenient point on the line. A shorting bar is then moved along the line. Two positions of the shorting bar that give the same reading of the indicator are one-half wavelength apart. In air the wave velocity on the line is essentially the same as that in free space. Hence,

$$f = \frac{150}{\Delta l} \qquad \text{Mc} \tag{5}$$

where Δl is in meters. The line should preferably be several half wavelengths long so that its resonant impedance will be essentially the same at the two settings of the shorting bar.

A second technique for using Lecher wires consists in moving the resonance indicator instead of the shorting bar. The resonance indicator may consist of a rectifier-microammeter and a very small antenna. With this device, the field near the transmission line can be explored. The distance between two successive maximums or minimums again equals $\lambda/2$. It is better to use two positions not closer than $\lambda/2$ or so to the shorting bar because the field is distorted near the bar. Thus, the distance from the bar to the nearest voltage minimum will be less than $\lambda/2$ because of the "end effect." The distance between the first and second minimums, however, is practically unaffected. This distance will be affected, however, by the presence of dielectric material. Hence, the portion of the line reserved for measurements should be as free as possible of insulators, supports, and other dielectric or conducting materials.

Strong harmonics sometimes cause difficulty on Lecher-wire systems. In extreme cases it may be necessary to plot a curve of the voltage distribution on the line to determine the lowest-frequency mode.

At the higher frequencies, a slotted coaxial line makes a satisfactory resonant system. A probe, inserted through the slotted portion, serves to explore the field along the line. Otherwise, the technique is substantially the same as for Lecher wires.

18. Cavity Wavemeters. At microwave frequencies the concept of lumped-parameter circuits has little meaning and resonant cavities are functionally equivalent to the tuned circuits used at lower frequencies. They are extensively used as wavemeters. They are mechanically simple and are capable of surprisingly high accuracy. Typical

constructions are sketched in Fig. 21.[7] The cavity takes the form of a cylinder tuned by a piston as in Fig. 21*a* or a plunger as in Fig. 21*b*. Tuning is accomplished by means of a micrometer screw mechanism which permits very good resolution. Such cavity resonators may have many resonant modes, and this can lead to ambiguities and parasitic resonances. A discussion of these problems is too involved to be presented here, but through careful design, difficulties arising from these sources can be largely eliminated. See Chap. 6, Sec. 48.

FIG. 21. Typical forms of cavity wavemeter.

In use, the cavity wavemeter is loosely coupled to the source by means of a loop or orifice. Resonance is commonly determined by noting the effect of the energy absorbed by the wavemeter. It is also possible to provide the wavemeter with a rectifier-microammeter or other detector to serve as a resonance indicator.

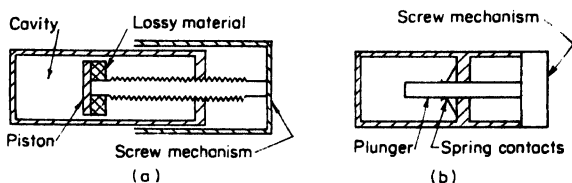
A well-constructed cavity wavemeter easily permits an accuracy of the order of 1 part in 1,000 to be obtained and with special care 1 part in 10^6 can be achieved. For such accuracy it is necessary to take into account second-order effects such as the influence of moisture vapor in the cavity and temperature.

19. Wave Form. Measurements of wave form can be approached from two points of view. When the details of the actual wave shape are the primary interest, a cathode-ray oscilloscope is indispensable. In this case wave forms are described in terms of rise time, pulse width, overshoot, and so on. When information of a more general kind is desired a harmonic wave analyzer is the appropriate instrument and wave forms are described in terms of harmonic content, distortion, and similar ideas.

A general-purpose oscilloscope is provided with amplifiers for both vertical and horizontal deflection, with sweep-voltage generators and synchronizing circuits as well as with the power supply necessary for the cathode-ray tube itself. In most cases the versatility of the instrument is limited primarily by the band width and gain of the amplifiers and the flexibility of the sweep-voltage generator. The useful band width of the CRT itself is of the order of 500 to 1,000 Mc and it can be made even higher. However, conventional video amplifiers provide a band width of up to only about 10 Mc. Through the use of distributed amplifiers this limit can be increased to 200 Mc or higher.

The sweep-voltage generator provides a saw-tooth voltage, *i.e.*, a voltage that increases linearly with time and then decreases very rapidly to its initial value. During the retrace time, a blanking voltage is applied to the CRT so that the beam is extinguished while the sweep voltage returns to begin a new trace. Thus, the horizontal position of the beam is proportional to time, and consequently, the pattern displayed on the screen shows the vertical-axis voltage as a function of time. Other voltage waves are sometimes used for the sweep voltage but the saw tooth is ordinarily most useful.

The sweep generator is provided with circuits that permit its frequency to be synchronized with that of the wave to be studied or some subharmonic of it. In this way stationary patterns are displayed on the screen.

When the wave to be studied is not periodic or the interesting part represents only a portion of a cycle, a so-called *driven sweep* is preferred. The driven sweep generates a saw-tooth voltage which is initiated by a trigger impulse usually derived from the voltage under study. A delay line in the vertical amplifier delays this voltage until the trigger has had sufficient time to start the sweep cycle. In this way the beginning of the wave is presented. Without the delay line, the beginning of the wave would be lost. Of course, trigger pulses may also be obtained from a signal related to the one under observation if desired, for example, a subharmonic.

Cathode-ray oscilloscopes are also used to provide a *parametric display*. For example the x-axis voltage may be the input voltage to an amplifier while the y-axis voltage is its output. If the horizontal and vertical amplifiers in the CRO are identical, and the output voltage of the amplifier is proportional to its input voltage, a straight line will result. Deviations from a straight line may represent distortion, phase shift, or other characteristics of the amplifier. When used in this way it is particularly important that the horizontal and vertical amplifiers of the CRO have similar characteristics. Otherwise, their differences may be attributed to the circuit under study.

20. Pulse Shape. The terms used in describing pulse shape are essentially descriptive. Quantities of chief interest are *amplitude, width, rise time, fall time, overshoot,* and *sag*. The significance of these terms is illustrated in Fig. 22.

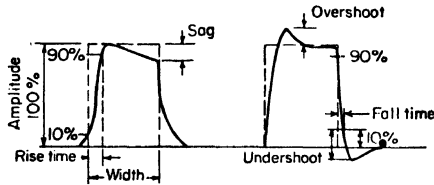



FIG. 22. Pulse-shape parameters.

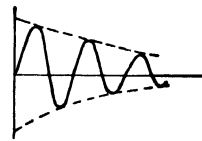



FIG. 23. Oscillatory pulse.

It is apparent from the figure that they are best adapted to describing pulses that are approximately rectangular in shape. Attempts to give precise definitions to these terms are not very rewarding unless the class of pulses to which they are applied is severely restricted. For example, there would be little point in discussing the rise time of a highly oscillatory pulse such as that shown in Fig. 23 since this is not likely to be a significant parameter.

Rectangular pulses, however, are used for many applications, especially for testing amplifiers, delay lines, etc. When a rectangular pulse is applied to an amplifier, the rise time of the output pulse is approximately inversely proportional to the band width of the amplifier, provided the rise time of the applied pulse is short compared with that of the output pulse. Amplifier circuits providing short rise times are likely to introduce overshoot. Consequently, rise time and overshoot are the parameters of most frequent interest.

The exact time at which a pulse begins or ends is rather difficult to specify. Hence, rise time is commonly specified as the time required for the pulse to increase from 10 to 90 per cent of its final value. Similarly, overshoot is usually expressed as a percentage of the final value. Other useful definitions are indicated in the figure.

21. Pulse Measurements. Pulse measurements are usually made with a wide-band oscilloscope that has a horizontal sweep generator calibrated in microseconds per centimeter. Provisions for calibrating the gain of the vertical amplifiers are also desirable. The rise time and pulse width can be determined by expressing these quantities in centimeters and then multiplying by the scale factor as shown by the dial settings of the oscilloscope. A convenient way to check the sweep speed calibration is to measure the period of a sine wave or square wave of known frequency.

It is important that the oscilloscope amplifiers be reasonably linear and care must be taken not to overload them. Obviously, an overloaded amplifier can clip a pulse, making it appear flat-topped when it is not, or otherwise distort the pulse. In addition, the rise time of the oscilloscope amplifiers must be small compared with the rise time of the observed pulse. An approximate formula for the total rise time of several amplifiers in cascade is

$$T \cong \sqrt{T_1^2 + T_2^2 + \cdots + T_n^2} \tag{6}$$

where $T_1, T_2, \cdots, T_n$ are the rise times of the respective amplifiers. This formula is accurate to a few per cent if none of the amplifiers has appreciable overshoot in its response.

When the band width of an amplifier is known, its rise time can be computed approximately from the relation

$$T \approx \frac{0.35}{W} \tag{7}$$

where T is the rise time in microseconds and W is the amplifier band width in megacycles. The half-power band width of the amplifier may be used for W, provided its pulse response is not oscillatory. As has been indicated above, there is often considerable uncertainty about the precise meaning of rise time and, therefore, these approximate formulas are sufficiently accurate for most practical purposes. When more accurate results are needed it is generally necessary to take into account the requirements of the particular problem under study as well as the detailed characteristics of the instruments.

When dealing with long pulses, for example, as with a low-frequency square wave, inadequate low-frequency response of the amplifier may introduce sag. When the total sag is of the order of 10 per cent or less, the total per cent sag is approximately the sum of the per cent sags introduced by each amplifier.

22. Wave Analyzers. For many purposes the wave shape of a signal is of less importance than its Fourier spectrum. If it is possible to obtain a photograph of the wave (a number of cameras are available which can be used to photograph the oscilloscope presentation) a graphical analysis can be used to determine the spectrum. However, it is often more convenient to use a wave analyzer. Although a wave analyzer does not yield information concerning the relative phase of a harmonic the phase is often of no importance.

A wave analyzer is essentially a very selective (narrow-band) calibrated amplifier combined with VTVM. The amplifier is tuned to the signal in question by adjusting for maximum deflection of the VTVM which then indicates the amplitude of the signal. Wave analyzers commonly employ the superheterodyne principle to achieve very high selectivity. One commercial wave analyzer has an effective band width of only 4 cps.[8] Others have adjustable band widths.

If a "true" Fourier spectrum is desired, the wave analyzer should have a very narrow pass band. However, the narrow band is a disadvantage in some cases. If the signal to be measured is drifting slightly or if its frequency fluctuates, as is often the case, the narrow-band analyzer may give misleading results. For example, suppose an a-f oscillator has its frequency modulated slightly by 120-cps power-supply ripple. The narrow-band analyzer would distinguish between the side bands produced by the modulation (these would be spaced at 120-cps intervals on each side of the average frequency) whereas the wide-band analyzer would indicate the average total voltage. Again, the narrow-band analyzer must have a slow-acting VTVM and it is easy to miss small signals by tuning past them before the meter can register. The wide-band analyzer is more tolerant in this situation. Thus, each type of instrument has its own field of application.

One of the chief limitations of the heterodyne wave analyzer is cross modulation which can occur in the mixer or preamplifier. With careful design spurious readings can be kept 60 db below the largest component at the input but difficulty may arise when an attempt is made to measure a small signal in the presence of a much larger one.

23. Phase Measurements. It is possible to discuss the phase difference of any two periodic signals having the same period and essentially the same wave form. Usually, however, the interpretation of phase difference is difficult except for wave forms that are sinusoidal or nearly so. For other wave shapes, the terminology of pulse techniques, *i.e.*, time delay, is likely to be more useful. Also, it is convenient to speak loosely of the relative phase of two sinusoidal waves of slightly different frequency. For example, the beat phenomenon associated with two waves of approximately the same frequency is often described by saying the waves are alternately in phase and out of phase. In its strict sense, phase difference refers to sinusoidal waves of the same

frequency, and in the following discussion of phase-measurement techniques it is assumed that the signals in question are sinusoidal and have the same frequency.

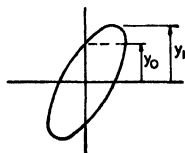
One of the easiest ways to measure phase difference at audio and low radio frequencies is by means of an oscilloscope. The two signals are applied to the horizontal and vertical deflection circuits, respectively. The resulting Lissajous figure then takes the form shown in Fig. 24. The phase difference θ between the two waves is given by the formula

$$\theta = \sin^{-1}\left(\pm \frac{y_1}{y_0}\right) \qquad (8)$$

Fig. 24. Lissajous figure for measuring phase difference.

If the phase shift introduced by the horizontal and vertical amplifiers of the oscilloscope is the same, the net phase shift indicated is the desired angle. If not, a correction must be made for the phase shift introduced by the oscilloscope. This can be determined by connecting one signal to both horizontal and vertical inputs. Then the angle can be measured and added to or subtracted from the indicated angle as required. It is desirable that the oscilloscope introduce very little phase shift. It is also clear that the method is suited only for waves that are very nearly sinusoidal.

The ambiguity concerning the proper quadrant for θ can be resolved by adding a small additional known phase shift to one of the signals. Sometimes this can be accomplished by changing slightly some parameter of the circuits under test. In other cases it may be necessary to adjust the test frequency to some value for which the phase shift is known approximately and then noting the direction in which the phase changes as the frequency is moved to the desired value. This, or something equivalent, is about the only way to resolve the whole-cycle ($360N$ deg) ambiguity that always is present. In most cases, it is possible to estimate the phase to within 360 deg; in others, the whole-cycle ambiguity is not objectionable.

If a calibrated phase shifter is available it can be used to introduce a known phase shift in one of the signals. The oscilloscope can then be used to determine when the net phase shift is zero and thus that the added phase is equal to the actual phase difference.

When a calibrated phase shifter is available, a variety of schemes can be used. For example, the two voltages whose phases are to be compared can be applied to a difference amplifier. The amplitude and phase are then adjusted by means of the phase shifter and an attenuator applied to one of the signals. When the output of the difference amplifier is zero, the phase difference can be read from the phase shifter. Care must be taken to be certain that the attenuator does not introduce appreciable phase shift. Also, as in any other measurement, precautions are necessary to assure that the phase-measuring equipment does not itself affect the phase difference to be measured.

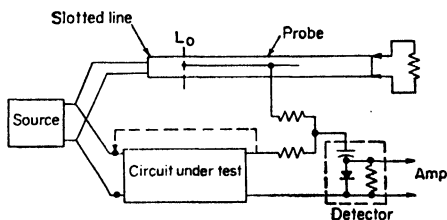



Fig. 25. Phase comparison with slotted line.

In the microwave region, a matched transmission line or slotted line may be used as a phase shifter. The general arrangement is shown in Fig. 25. The line should be matched to a standing-wave ratio of 5 per cent or better to avoid corrections to compensate for the mismatch. If the source is modulated at an audio frequency the detector may consist of a crystal rectifier and audio amplifier. The system is "zeroed" by connecting the differencing circuit to the source side of the two circuits and then adjusting the probe for minimum detector output. The two signals at the detector input should be approximately the same amplitude so that a well-defined minimum will be obtained. If a suitable attenuator is available a null may be produced by

adjusting the signals to be equal in amplitude and 180 deg out of phase. (If a difference amplifier is used as a detector, the null occurs when the two signals are in phase.) Two successive nulls correspond to probe positions separated by one wavelength or 360 deg.

24. Direct-reading Phase Meter. Several schemes have been devised to provide direct-reading phase meters. With the help of pulse techniques, the signals to be compared are converted to square waves or pulses of equal amplitude. In one system, positive pulses from the two signals are applied to the grids of a flip-flop circuit. This flip-flop is switched to the "on" state by pulses derived from one signal and to the "off" state by pulses from the other signal. Consequently, the fraction of a cycle that one tube is on is proportional to the phase angle between the input signals as indicated in Fig. 26. Hence the average current as indicated on a d-c meter is also proportional to this angle. The current in the other tube of the flip flop is proportional to 360 deg $-$ θ.

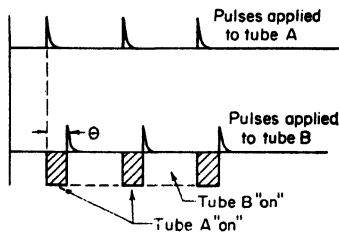



Fig. 26. Pulse-angle relationships in direct-reading phase meter.

The pulses are generated so that they occur when their associated signals change sign from negative to positive, *i.e.*, at the positive-going zero crossings. Continuous readings from zero to 360 deg are obtained.

The polarity of either signal can be reversed; thus, 180 deg $-$ θ can also be measured. This instrument is not sensitive to wave-form errors provided the distortion does not cause the zero crossings to be unequally spaced and does not cause more than two zero crossings per cycle.

The useful frequency range of such a device is limited by the switching times of the pulse generators and flip flops. In one commercial instrument an accuracy of about 5 deg is obtained from 10 cps to 40 kc. The range can be extended upward through the use of heterodyne techniques since the phase angle between two signals can be preserved if they are heterodyned with the same signal.

POWER MEASUREMENT

Although several instruments have been devised for the direct measurement of power at high frequencies in terms of current, voltage, and power factor, they have not found widespread use. The reason probably is that under the conditions for which high-frequency power measurements are usually made indirect methods are both satisfactory and more practical.

25. Methods Employing Standard Impedances. For measuring the power output of an a-f or r-f amplifier the simplest method is to dissipate the power in a standard resistance and measure the voltage with a VTVM or the current with a thermocouple ammeter. The power is then computed from E^2/R or I^2R.

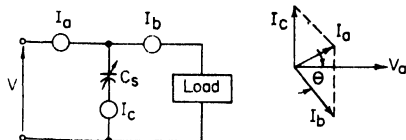



Fig. 27. Three-ammeter method for measuring power.

When a standard resistance of suitable value is not available the impedance of the circuit used can be measured at the operating frequency by means of a bridge and the power calculated from either E^2/R_p or I^2R_s, in which R_p is the equivalent parallel resistance and R_s is the equivalent series resistance of the circuit in question.

The same principle can be used to measure the power input to an antenna, matched transmission line, or other circuit whose impedance is known.

A variation of the principle known as the *three-ammeter method* is sometimes useful. This method also requires a known impedance. The circuit arrangement is shown in

Fig. 27. Neglecting the losses in C_s

$$P = 2X_c \sqrt{S(S - I_a)(S - I_b)(S - I_c)} \tag{9}$$

in which P = power, watts
X_c = reactance of C_s, ohms
$S = \frac{1}{2}(I_a + I_b + I_c)$

A similar scheme is shown in Fig. 28. In this case

$$P = \frac{1}{2R} \sqrt{V_a^2 - V_b^2 - V_c^2} \tag{10}$$

If the resistor is replaced by a capacitor as in Fig. 29, then

$$P = \frac{2}{X_c} \sqrt{S(S - V_a)(S - V_b)(S - V_c)} \tag{11}$$

in which $$S = \frac{V_a + V_b + V_c}{2}$$

It is clear from the vector diagrams that best accuracy is obtained when the three currents (or voltages) are the same order of magnitude and the currents are nearly sinusoidal. Harmonic currents in the capacitance in Fig. 27 can introduce large errors even though the voltage across the load may have a good wave form.

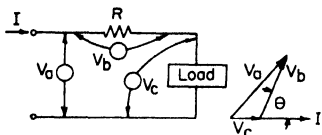



Fig. 28. Three-voltmeter method for measuring power.

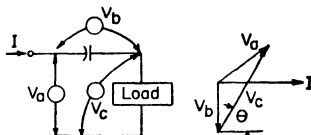



Fig. 29. Variation of three-voltmeter method.

26. Bolometer Methods. The methods described in Sec. 25 are useful at frequencies up to about 500 Mc. Above this frequency it becomes increasingly difficult to measure current and voltage. At these higher frequencies and at microwave frequencies energy is nearly always distributed over matched transmission lines or wave guides. When the amount of power to be measured is relatively small it is dissipated in a specially constructed resistor called a bolometer. The term bolometer is used to refer to either of two kinds of temperature-sensitive resistors: *barretters*, which are metallic resistors with positive temperature coefficients of resistance, and *thermistors* made of metallic oxides and which have negative temperature coefficients of resistance. Barretters are usually less rugged than thermistors but since they operate at higher temperatures they are less affected by ambient temperature.

Bolometers can be used to measure power of less than 1 μw. When relatively large amounts of power are to be measured, a bolometer can be used to "sample" the power, *i.e.*, to measure a small amount of power that is proportional to the desired power.

Two methods are in common use to measure power by means of bolometers.[7] In each case the bolometer forms one arm of a Wheatstone bridge. A *balanced-bridge* arrangement is shown schematically in Fig. 30. To measure the r-f power, the a-f power supply is turned off, and the bridge is balanced by adjusting the d-c bias current. Then the r.f. is turned off and the bridge again balanced by adjusting the a-f voltage. Now the a-f power supplied to the bolometer exactly equals the power previously supplied by the r-f source. Since the bridge is balanced, this power equal $V^2/4R$.

The *unbalanced-bridge* method is similar, but the a-f source is not required. Instead, the bridge is balanced by adjusting the bias current with the r-f power turned off.

When the r.f. is turned on the bridge is, of course, unbalanced again. The null detector is calibrated to indicate directly the r-f power. This method is more convenient than the balanced-bridge method, but the change in resistance of the bolometer during the measurement is not always permissible.

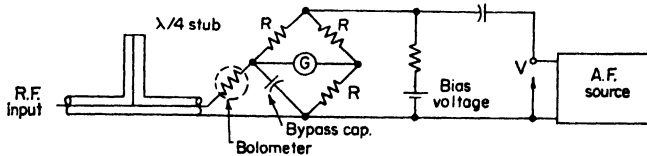



FIG. 30. Bolometer bridge.

27. Directional Couplers. A directional coupler can distinguish between the waves traveling in opposite directions when connected to a wave guide or transmission line. It can be arranged to respond to either the incident or the reflected wave or to both. When arranged to respond to both waves the signal associated with one wave appears at one output, while that associated with the other wave appears at a different output.

A bolometer connected to one terminal of a directional coupler affords a convenient way to monitor the power transmitted over the system. The power required by a properly designed bolometer is so small that it has a negligible effect on the system. If desired, the transmitted and reflected power can be measured separately and the actual load power computed as their difference. However, the reflected power is relatively small even for rather poor matching. For example, a voltage standing-wave ratio of 2 corresponds to a reflection coefficient of $\frac{1}{3}$ and the reflected power is only $\frac{1}{9}$ the load power. For a voltage standing-wave ratio of 1.25 the reflected power is only a little more than 1 per cent of the load power. Thus, it is evident that it is usually sufficient to measure only the transmitted wave.

Voltage measurements can also be used to sample the power in a matched transmission line. However, unless the line is perfectly matched, the voltage reading depends on the location of the probe relative to the standing-wave pattern, whereas the directional coupler is not affected by the standing waves. The superiority of the directional coupler is illustrated by the above examples.

CIRCUIT PARAMETERS

Three general methods are available for measuring circuit parameters at high frequencies. These are the voltmeter-ammeter method, the substitution method, and the bridge method. In addition, a number of direct-reading impedance meters are commercially available. The voltmeter-ammeter method is not well suited to high-frequency applications because sensitive a-c milliammeters are not available and the angle of an impedance is usually desired as well as its magnitude. Consequently, measurements in terms of voltage, current, and power are cumbersome. Substitution measurements are easy to make but have the disadvantage that an adjustable standard having the same impedance as the unknown is required.

Bridge measurements afford the best accuracy. However, when a bridge suitable at the required frequency is not available satisfactory results can be obtained in other ways. With a good vacuum-tube voltmeter, a standard variable capacitor, and a Q meter, methods can be improvised to meet a surprisingly wide range of needs.

28. Equivalent Circuits. For some purposes it is desired only to measure the equivalent resistance and reactance of a circuit or component at a particular frequency. In other cases it is more useful to determine an equivalent circuit that is valid over a substantial frequency band. For this purpose it is necessary to make measurements at more than one frequency and it is advisable to make additional measurements over the band width of interest.

When making measurement on coils it is always wise to estimate the effects of the distributed capacitance even when its value is not required. The equivalent circuit in

Fig. 31*a* is a good representation for coils of usual design. Analysis of this circuit shows that the "apparent" inductance, *i.e.*, the equivalent inductance, is approximately

$$L_e \approx L/(1 - \omega^2 L C_0) \tag{12}$$

or

$$L_e \approx \frac{L}{1 - (\omega/\omega_0)^2} \tag{13}$$

in which L is the true inductance, ω is the frequency at which the measurement is made, and ω_0 is the natural resonant frequency of the coil.

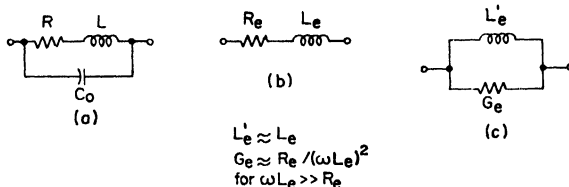



Fig. 31. Equivalent circuits for coil with distributed capacitance.

The equivalent resistance is given by

$$R_e \approx \frac{R}{[1 - (\omega/\omega_0)^2]^2} \tag{14}$$

Equations (12) to (14) are obtained by assuming that $R\omega C_0 \ll (1 - \omega^2 L C_0)$, which is equivalent to assuming $1 \ll \frac{|X_c| - |X_L|}{R}$. Obviously, they cannot be used if the operating frequency is near the natural frequency of the coil. However, if the Q of the coil at its natural frequency is greater than about 10, these equations will give meaningful results for frequencies less than about 70 per cent of the natural frequency.

Figure 32 shows how the impedance of the circuit in Fig. 31*a* varies with frequency. It is clear that the shunt capacitance C_0 can be neglected only at frequencies that are *well below* the natural frequency.

These considerations are significant when measuring coil characteristics. Thus, if the inductance and resistance (or Q) of a coil are measured on a bridge or conventional Q meter the results may be quite misleading if the coil is subsequently used as part of a parallel-resonant circuit. Therefore, the choice of measurements to be made will depend on the application involved.

Fig. 32. Variation of impedance with frequency for circuit in Fig. 31*a*.

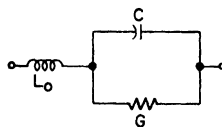



Fig. 33. Equivalent circuit for capacitor with residual inductance.

Analogous but less troublesome difficulties are caused by the residual inductance of capacitors as indicated in the equivalent circuit of Fig. 33. The residual inductance L_0 is seldom as large as the inductance of connecting wires and consequently it can be neglected except for very exacting measurements. However, certain types of rolled-paper capacitors have an appreciable inductance and cannot be used at high frequencies.

29. Substitution Measurements. Substitution measurements are easy to make and quite satisfactory when suitable standards are available. Basically, the method consists of measuring some characteristic of a circuit that includes the unknown to be

determined and then substituting a standard element for the unknown. When the standard has been adjusted to duplicate the original conditions, the standard has the same value as the unknown. A feature of the method is that it depends on the accuracy of the standard rather than on the accuracy of an instrument such as a voltmeter or ammeter.

The method is illustrated in Fig. 34. In (*a*) an ammeter is connected to a low-impedance source E in series with the unknown R_x. After the current is noted, the standard is connected in place of the unknown and adjusted to produce the same current. Then $R_s = R_x$. The impedance of the source is not important provided it is low compared with R_x and does not change during the measurement. In (*b*) a high-impedance source is used or a resistance R is connected in series with a low-impedance source. In this case R_s is adjusted so that V is the same when R_s is substituted for R_x. Note that a low-impedance source should not be used in (*b*) because V would then depend very little on R_x.

FIG. 34. (*a*) Substitution method using current balance; (*b*) substitution method using voltage balance.

This method can be used to measure a capacitance or an inductance. However, if the resistance of the inductor is appreciable this method is unsatisfactory because it provides no way to determine the phase angle.

A convenient way to measure a capacitance, especially a small one, is to use it as part of the tuning capacitance in a resonant circuit. Several methods of coupling the tuned circuit to a signal generator are indicated in Fig. 35. The procedure is to adjust the circuit to resonance by varying the standard capacitor to obtain maximum voltage as indicated by the VTVM. The unknown capacitor is then disconnected and the circuit readjusted for resonance. The difference between the two settings of C_s is the value of C_x.

FIG. 35. Methods for coupling a tuned circuit to a signal generator.

This procedure can also be used to measure a large inductance provided the resonant frequency is known. In this case, the susceptance of the inductance equals the susceptance added by C_s when the circuit is returned. Since the inductive susceptance is negative, the capacitance must be *decreased* when the inductor is removed. Hence,

$$\frac{1}{\omega L_e} = \omega \, \Delta C_s$$

or

$$L_e = \frac{1}{\omega^2 \, \Delta C_s} \tag{15}$$

If it should happen that C_s must be increased when the inductor is removed, this indicates that the operating frequency is above the natural resonant frequency of the inductor. Note also that this method yields the equivalent parallel inductance. If the Q of the inductor is greater than 10, the equivalent parallel inductance and the equivalent series inductance differ by less than ½ per cent.

A variation of this procedure is to connect the unknown reactance in series with the standard capacitor. The change in the reactance of the standard then equals the unknown reactance. For a capacitance,

$$C_x = \frac{C_{s1} C_{s2}}{C_{s2} - C_{s1}} \tag{16}$$

and for an inductance

$$L_x = \frac{1}{\omega^2}\frac{C_{s1} - C_{s2}}{C_{s1}C_{s2}} \tag{17}$$

in which C_{s2} and C_{s1} are the values of the standard with and without the unknown, respectively.

30. Susceptance Variation. One of the most useful techniques for determining the equivalent parallel circuit of a component is the susceptance variation method and its related schemes. The circuit arrangement is shown in Fig. 36*a*. It is simpler to

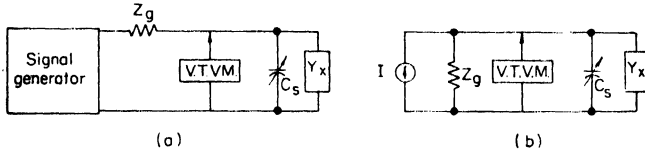



FIG. 36. (*a*) Circuit for susceptance-variation measurements; (*b*) equivalent circuit.

regard the unknown as an admittance consisting of a conductance G_x and a susceptance B_x in parallel. Z_g should be an impedance having a magnitude as large as or larger than the impedance across which the VTVM is connected. A resistance is best but a small capacitor can be used.

One of the chief advantages of this circuit is that each component except Z_g can have one of its terminals connected to a common ground.

Applying Norton's theorem to the circuit in Fig. 36*a* gives the circuit in Fig. 36*b*. Combining the input admittance of the VTVM with that of Z_g gives the circuit in Fig. 37, in which the voltmeter has zero input admittance. Assuming Y_x is inductive, the total conductance of the circuit is measured as follows. First, C_s is adjusted for resonance (maximum voltage) at a value C_1 and the voltage V_0 noted. Then C_s is increased to a value C_2 such that the voltage decreases to a new value V_1. Then the total conductance is

$$G_t = \frac{\omega(C_2 - C_1)}{\sqrt{(V_0/V_1)^2 - 1}} \tag{18}$$

and

$$G_x = G_t - G_0 \tag{19}$$

If Z_g is a resistance R_0 then G_0 is very nearly equal to $1/R_0$, since the losses in the capacitor are usually small enough to be neglected. If Z_g is a capacitance, C_0 can usually be neglected.

The susceptance of Y_x equals the total capacitive susceptance at resonance. Thus,

$$L = \frac{1}{\omega^2(C_1 + C_0)} \tag{20}$$

It is often sufficient to estimate the stray capacitance C_0 or to neglect it entirely.

If a coil whose susceptance and conductance are known is available, then Eqs. (19) and (20) can be used to find G_0 and C_0 and thus to improve the accuracy of the measurements.

The parameters of a known coil can be determined as follows. First, the total conductance of the circuit as in Fig. 37 is determined by Eq. (18). Then the known coil is added to the circuit (care being taken to assure that there is no mutual inductance between the two coils). Now the total conductance is again measured and the increase is the conductance of the known coil. Its susceptance is the change in susceptance ($\omega\,\Delta C_s$) required to retune the circuit when the known coil is added. Finally, the known coil is substituted for Y_x in Fig. 37 and G_0 determined. Also, C_0 can be found since at resonance

$$\omega(C_s + C_0) = \frac{1}{\omega L_x} \tag{21}$$

and C_s is known.

To measure the capacitance and conductance of a capacitor, they are simply added to the circuit of Fig. 37 with a coil for Y_x. The increase in total conductance is the conductance of the unknown and its capacitance is the change in C_s necessary to retune the circuit when C_x is added. Neither C_0 nor G_0 need be known for this measurement.

FIG. 37. Circuit equivalent to Fig. 36b.

Near resonance the voltage changes slowly with C_s so the value corresponding exactly to resonance is not well defined. The accuracy can be improved by finding two values C_2 and C_3, one greater than the value for resonance and the other less than the value for resonance and for which the voltages are equal. Then

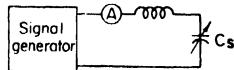
$$C_1 = \frac{C_2 + C_3}{2} \tag{22}$$

Using this value for C_1 in Eq. (18) gives

$$G_x = \frac{(\omega/2)(C_2 - C_3)}{\sqrt{(V_0/V_1)^2 - 1}} \tag{23}$$

The value V_0/V_1 is not critical. A convenient value is $\sqrt{2}$ since the radical in the denominator of Eq. (23) then equals unity. However, when the conductance is very small a larger value may improve the accuracy by increasing $(C_2 - C_3)$. On the other hand, if too large a value is chosen the voltage variation again becomes small and the accuracy suffers.

31. Reactance Variation. Instead of the parallel circuit a series arrangement may be used as indicated in Fig. 38. The same principles apply to this circuit as to Fig. 37 but in this case the current in the circuit is observed. Results are expressed in terms of the equivalent series resistance and reactance. The susceptance variation method described in Sec. 30 is generally more convenient to use at radio frequencies.

FIG. 38. Circuit for reactance-variation measurements.

32. Frequency Variation. In modified versions of both the preceding methods, the frequency can be varied instead of the capacitance. Greater care is required when frequency is varied because the values of the elements in the equivalent circuits may depend upon frequency. For narrow-band circuits, *i.e.*, circuits for which Q is large, the frequency-variation method gives good results. The signal generator must be arranged so that small increments can be read with good accuracy. The basic measurement is the half-power band width of the tuned circuit. Ideally, a current source should be used to excite the tuned circuit but the arrangement in Fig. 39 is satisfactory. When the output voltage of the signal generator is held constant this circuit is equivalent to a current source driving the unknown circuit in parallel with R_0. The conductance of R_0 should be subtracted from the measured total.

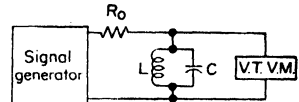



FIG. 39. Circuit for frequency-variation measurements.

The procedure for measuring the total conductance G_t is to adjust the frequency for resonance (maximum voltage) and note ω_0 and V_0. Next, adjust the frequency to a new value ω_1 which gives a reading V_1. Then

$$G_t = \frac{2C(\omega_1 - \omega_0)}{\sqrt{(V_0/V_1)^2 - 1}}$$

The approximation is within ½ per cent when Q is greater than 10.

An exact relation is obtained by using two frequencies for which the voltage readings are both V_1. Assume ω_2 is less than ω_0; then

$$G_t = \frac{C(\omega_1 - \omega_2)}{\sqrt{(V_0/V_1)^2 - 1}} \tag{24}$$

With $(V_0/V_1)^2 = \sqrt{2}$, this yields the very simple relation

$$G_t = C(\omega_1 - \omega_2) \tag{25}$$

Also,

$$G_x = G_t - G_0 \tag{26}$$

This scheme is well suited to measuring the Q or the band width of a tuned circuit since it avoids the necessity of making separate measurements on L and C. The impedance at resonance is a pure resistance,

$$Z_R = 1/G_x \tag{27}$$

and the circuit Q is

$$Q_e = \frac{\omega_0}{\omega_1 - \omega_2} = \frac{f_0}{f_1 - f_2} \tag{28}$$

in which f_1 and f_2 are the half-power frequencies for which $V_0/V_1 = \sqrt{2}$. Other useful relations are

$$C = \frac{G_x}{\omega_1 - \omega_2} \tag{29}$$

$$L = \frac{\omega_1 - \omega_2}{\omega_0{}^2 G_x} = \frac{1}{Q\omega_0 G_x} \tag{30}$$

and

$$\omega_1\omega_2 = \omega_0{}^2 \tag{31}$$

As determined by this method, G_x, C, and L are the values in the equivalent parallel circuit. Thus, G_x includes the losses in the capacitor and C includes the distributed capacitance of the coil.

The frequency-variation technique can also be used for measurement on individual components by noting the changes they produce in ω_0 and G. It must not be forgotten that the values in the reference equivalent circuit will depend on the frequency to some extent.

33. Distributed Capacitance of Coils. It is possible in principle to determine the distributed capacitance of a coil by measuring the resonant frequencies when two known values of capacitance are added and using the relationships

$$\omega_1{}^2 = 1/L(C_1 + C_0)$$
$$\omega_2{}^2 = 1/L(C_2 + C_0)$$

to solve for C_0. However, better accuracy can be obtained by plotting $1/f^2$ as a function of the added capacitance as indicated in Fig. 40. If the resulting straight line is then extrapolated, the X-axis intercept gives the distributed capacitance C_0 as in Fig. 40.

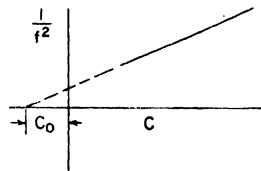



FIG. 40. Graphical determination of distributed capacitance.

It should be remembered that the capacitance of a coil is in reality distributed and the lumped-circuit approximation is not suitable for every purpose. Above the lowest parallel-resonant frequency most coils behave like distributed-parameter systems exhibiting both series and parallel resonances. Since coils are seldom intentionally used in this frequency range this behavior is usually of practical importance only when it produces unwanted results. Then an equivalent circuit must be found to represent the coil over the range in question. It may also be added that the self-resonant frequency of a coil is usually somewhat higher than the value calculated from measurements of L and C_0 by the above methods.

34. Measurements of Q. The Q of a coil is defined by the relation

$$Q_L = \frac{\omega L}{R} \tag{32}$$

in which L is the inductance and R the resistance of the coil.

The Q of a resonant circuit is defined by the relation

$$Q_e = \frac{\omega_0}{W} = \frac{2\pi f_0}{2\pi W} = \frac{f_0}{W} \tag{33}$$

in which f_0 is the resonant frequency and W is the half-power band width. In the case of a parallel-resonant circuit the half-power band width is the difference between the two frequencies for which the voltage is $1/\sqrt{2}$ times the voltage at resonance when the circuit is excited by a constant-current source. In the case of a series-resonant circuit, the half-power band width is the difference between the two frequencies for which the current is $1/\sqrt{2}$ times the current at resonance when the circuit is excited by a constant-voltage source.

FIG. 41. Ideal parallel GCL circuit.

When the only losses in a resonant circuit are in the coil, the circuit Q is the same as the coil Q. Hence, the circuit Q is often referred to as the equivalent Q. In the idealized parallel circuit in Fig. 41 the circuit Q is given by

$$Q_e = \frac{\omega_0 C}{G} = \frac{1}{G\omega_0 L} \tag{34}$$

To find the Q of a soil, its inductance and resistance can be measured by any of the methods described in Secs. 33 to 39 and the Q can be computed from the appropriate one of Eqs. (32), (33), and (34). Equation (28) gives the circuit Q directly according to Eq. (33). When making calculations based on conductance measurements one obtains circuit Q or coil Q depending upon whether the total conductance or only the coil conductance is used. Thus, if the coil conductance is used in Eq. (34), the result is the same as if the equivalent series-circuit values are computed and used in Eq. (32). Equations (18) and (20) give, for the susceptance-variation method,

$$Q_e = \frac{C_1 + C_0}{C_2 - C_1}\sqrt{(V_0/V_1)^2 - 1} \tag{35}$$

Equation (35) emphasizes the necessity of considering the distributed capacitance when making Q measurements. However, the frequency-variation method yields correct circuit-Q values without the necessity of measuring C_0.

35. The Q Meter. The Q meter provides the most convenient and the most widely used method for measuring Q. A simplified circuit diagram of a Q meter is shown in

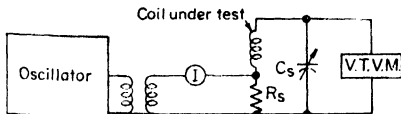



FIG. 42. Simplified circuit of Q meter. FIG. 43. Equivalent circuit of Q meter.

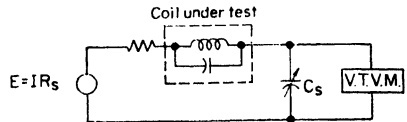



Fig. 42. In this device, a known voltage is introduced in series with the coil under test by passing a known current through a standard resistance. Notice that the tuned circuit is in series resonance when the capacitor is adjusted for maximum voltage. An equivalent circuit as shown in Fig. 43 can be obtained with the help of Thévenin's

theorem. At resonance the capacitor voltage is

$$V_c = \frac{E}{R_s + R_c}\frac{1}{\omega C} = E\frac{\omega L}{R_s + R_e} \tag{36}$$

or

$$V_c/E = Q_e \tag{37}$$

Since $E = I_0 R_s$, the VTVM can be calibrated to read Q directly when I_0 is set to predetermined values.

It is clear that the Q meter measures equivalent Q and not coil Q. The coil Q can be found by making corrections for the distributed capacitance of the coil and the resistance R_s. Of course, these corrections are not required for most routine measurements; hence, this feature of the Q meter is not a serious disadvantage. When C_0 and R_s are known the coil Q is readily computed:

$$Q_L = Q_0 \frac{C_s - C_0}{C_s}\frac{R_0}{R_0 - R_s} \tag{38}$$

In this expression Q_0 = observed Q
$R_0 = 1/\omega L_s Q_0$ = observed resistance

It is clear that the correction for C_0 may easily be quite large. The correction for R_s is seldom significant.

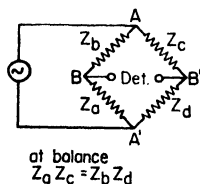
Another feature of the Q meter that is occasionally important arises from the fact that the maximum capacitor voltage does not occur exactly at series resonance (unity power factor). When the circuit Q is less than 10, the difference between the capacitances (or frequencies) producing maximum V_c and those for true resonance is appreciable.

The Q meter is also very useful for making substitution measurements and for measuring distributed capacitance of coils as described in Sec. 29. Commercial instruments are calibrated to measure inductance directly by setting the frequency to predetermined values. The direct reading is, of course, equivalent inductance which must be corrected for C_0 when circumstances demand it.

The Q meter furnishes a very practical way for making a wide variety of routine measurements.

BRIDGE MEASUREMENTS

As discussed earlier, impedance measurements in the frequency range below 150 Mc can be made with more precision by means of bridges than with any other method. Above this frequency stray capacitances and inductances cause serious errors in the measurements.

36. Wheatstone Bridge. The bridges used for measuring impedance at audio and radio frequencies have the general form of the Wheatstone bridge shown in Fig. 44.

When the bridge is balanced no potential difference exists between points B, B' and the detector output is a null. Under this condition

$$Z_a Z_c = Z_b Z_d \tag{39}$$

Fig. 44. General form of Wheatstone bridge.

The impedances in the arms are complex quantities of the form $Z = R + jX$; therefore, Eq. (39) represents a pair of equations in the real variables R and X:

$$R_a R_c - X_a X_c = R_b R_d - X_b X_d \tag{40a}$$
$$R_a X_c + R_c X_a = R_b X_d + R_d X_b \tag{40b}$$

Since the impedances making up the several arms of the bridge may be composed of various combinations of R, L, and C many variations of the Wheatstone bridge are used. Several of the more common forms together with the equations of balance are

shown in Figs. 45 and 46. Some of these are discussed later in this section. The general equations of balance, Eqs. (39) and (40), are valid for all of these.

37. Common Forms of Bridges and Their Uses. One of the most common forms of the Wheatstone bridge is the *resistance-ratio bridge*. Here (Fig. 45*a*) two adjacent arms, called the ratio arms, contain pure resistances. The impedances in the other two arms must have the same sign of reactance for the bridge to balance. When the

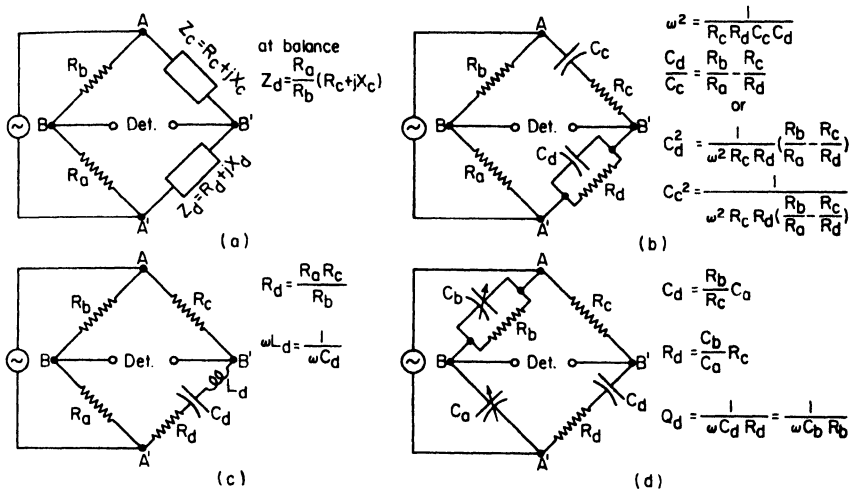



FIG. 45. (*a*) Resistance-ratio bridge; (*b*) Wien bridge; (*c*) resonance bridge; (*d*) Schering bridge.

bridge is balanced the phase angles of the two impedances are equal and their magnitudes are in the same ratio as the resistances in the ratio arms:

$$\frac{R_d}{X_d} = \frac{R_c}{X_c} \tag{41a}$$

$$\frac{|Z_d|}{|Z_c|} = \frac{R_a}{R_b} \tag{41b}$$

The balance condition can also be expressed as

$$\frac{R_d}{R_c} = \frac{R_a}{R_b} \tag{42a}$$

$$\frac{X_d}{X_c} = \frac{R_a}{R_b} \tag{42b}$$

This bridge is useful for measuring an inductance in terms of another inductance or a capacitance in terms of another capacitance but cannot be used to measure an inductance in terms of a capacitance in an adjacent arm or vice versa.

The *Wien bridge* (Fig. 45*b*) is a special case of the resistance-ratio bridge which permits two capacitances to be compared if all the resistances in the bridge are known. When the bridge is balanced

$$\omega^2 = \frac{1}{R_c R_d C_c C_d} \tag{43a}$$

and

$$\frac{C_d}{C_c} = \frac{R_b}{R_a} - \frac{R_c}{R_d} \tag{43b}$$

Both the capacitances can be computed from the resistances and the frequency if the frequency is also known. Solving Eqs. (43a) and (43b) for C_d and C_c yields

$$C_d^2 = \frac{1}{\omega^2 R_c R_d}\left(\frac{R_b}{R_a} - \frac{R_c}{R_d}\right) \tag{44a}$$

and

$$C_c^2 = \frac{1}{\omega^2 R_c R_d[(R_b/R_a) - (R_c/R_d)]} \tag{44b}$$

The Wien bridge also furnishes a convenient means of determining frequency in the a-f range. The frequency is measured in terms of resistance and capacitance as indicated in Eq. (43a). In a frequency meter of this type C_d is usually made equal to C_c and the two are "ganged" together so that the frequency at which a null occurs varies linearly with capacitance. This frequency meter, therefore, covers a much wider frequency range than does an LC meter employing the same capacitor since the frequency of the LC meter varies as the square root of the capacitance. It is also convenient to make R_d equal to R_c, in which case Eq. (43b) requires that

$$R_b = 2R_a$$

when Eq. (43a) reduces to

$$f = \frac{1}{2\pi R_d C_d} \tag{45}$$

R_c and R_d can be changed in increments to increase the frequency range.

The *resonance bridge* (Fig. 45c) is another form of the resistance-ratio bridge. Three of the arms are pure resistances; therefore Z_d must be real when the bridge is balanced. The balance equations are

$$R_d = \frac{R_a}{R_b} R_c \tag{46a}$$

$$\omega L_d = \frac{1}{\omega C_d} \tag{46b}$$

The resonance bridge is used to measure inductance in terms of capacitance or vice versa. It can also be used to measure frequency in terms of inductance and capacitance.

The *Schering bridge* (Fig. 45d) is used for measuring capacitance and resistance in terms of the same quantities. The balance equations are

$$C_d = \frac{R_b}{R_c} C_a \tag{47a}$$

and

$$R_d = \frac{C_b}{C_a} R_c \tag{47b}$$

Usually R_b and R_c are fixed and C_a and C_b are made variable. This permits both the capacitive and resistive components of the unknown impedance Z_d to be balanced out by varying capacitors. With R_b and R_c fixed, C_a can be calibrated to read directly the unknown capacitance. C_b can then be calibrated to read directly the Q or the dissipation factor $D = 1/Q$ of the unknown capacitor at one frequency:

$$Q_d = \frac{1}{\omega C_d R_d} = \frac{1}{\omega C_b R_b} \tag{48}$$

The Schering bridge, therefore, furnishes a convenient means for measuring the capacitance and dissipation factor of capacitors.

The bridges previously discussed are used for comparing two impedances with the same sign of reactance. Often, however, it is convenient to compare an impedance

with another having the opposite sign of reactance. For example, capacitors can be made with much higher quality factors than can coils and hence represent idealized circuit elements much more closely. Therefore, inductance is often measured on a bridge which employs a capacitor as the standard of comparison. When impedances with opposite signs of reactance are compared they must either be placed in the same arm of the bridge (resonance bridge) or in opposite bridge arms—never in adjacent arms.

One of the simplest bridges for comparing impedances with unlike reactances (Fig. 46a) places them in opposite arms with pure resistances in the other two arms.

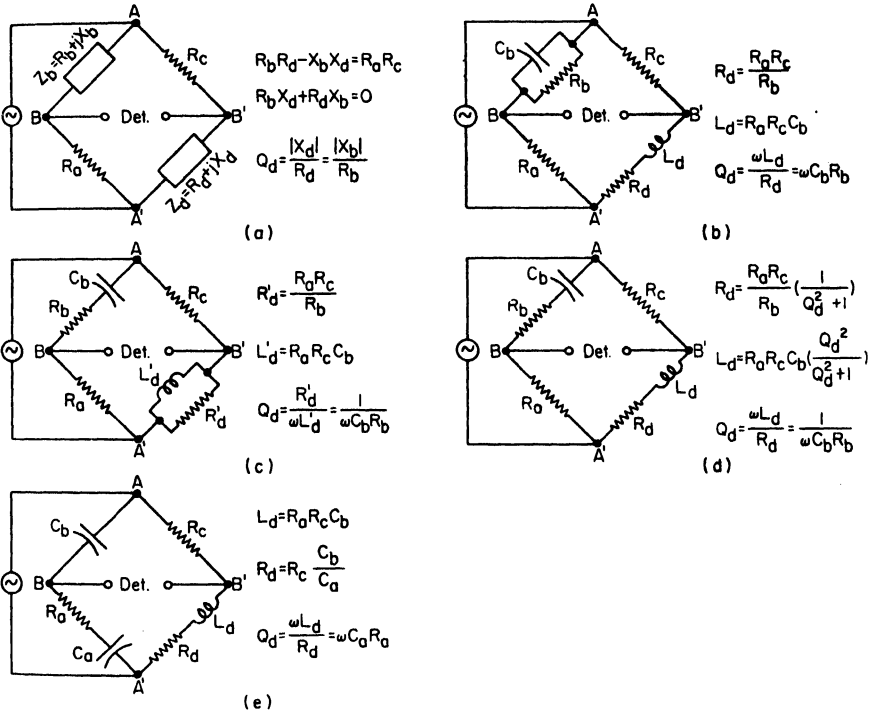



FIG. 46. (a) Bridge for comparing impedances with opposite signs of reactance; (b) Maxwell bridge; (c) Hay bridge with parallel inductance and resistance; (d) Hay bridge with series inductance and resistance; (e) Owen bridge.

Both the *Maxwell* and *Hay bridges* are of this form. This general form of bridge corresponds to the resistance-ratio bridge for comparing reactances with like signs. Since the product R_aR_c is real, Z_b and Z_d must have phase angles opposite in sign and of the same magnitude. The balance equations are

$$\frac{|X_d|}{R_d} = \frac{|X_b|}{R_b} \tag{49a}$$

and

$$|Z_d|\,|Z_b| = R_aR_b \tag{49b}$$

When neither Z_d nor Z_b contains both L and C, Eq. (49a) is equivalent to

$$Q_d = Q_b \tag{50}$$

The balance conditions can also be expressed as

$$R_bR_d - X_bX_d = R_aR_c \qquad (51a)$$
$$R_bX_d + R_dX_b = 0 \qquad (51b)$$

The Maxwell bridge (Fig. 46*b*) is used for measuring the inductance and Q of coils in terms of capacitance and resistance. The standard capacitor C_b is shunted with a resistance. The equations of balance are

$$R_d = \frac{R_aR_c}{R_b} \qquad (52a)$$

$$L_d = R_aR_cC_b \qquad (52b)$$

This bridge may be used in either of two ways. In one, a fixed standard capacitor C_b is used and the reactive balance is obtained by varying either R_a or R_c. The resistive balance is then obtained by adjusting R_b. The resistive balance is dependent on the reactive balance as can be seen from Eqs. (52) but the method has the advantage that R_a (or R_c) can be calibrated to read directly in inductance (independent of frequency) while R_b can be calibrated to read Q_d at any one frequency:

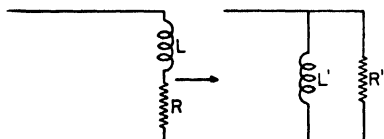
$$Q_d = \frac{\omega L_d}{R_d} = \omega C_b R_b \qquad (53)$$

$L' = L\left(\frac{Q^2+1}{Q^2}\right)$, $R' = R(Q^2+1)$, $Q = \frac{\omega L}{R} = \frac{R'}{\omega L'}$

Fig. 47. Approximate equivalent circuits of a coil far below its resonant frequency.

R_a and R_c are held constant in the second way of using the Maxwell bridge while C_d is made variable. This method has the advantage that the resistive and reactive balances are independent. C_b and R_b can now be calibrated so that the bridge reads directly in inductance and resistance. However, R_b cannot be calibrated to read Q_d.

It should be noted here that the representation of a coil by an inductance and a resistance is approximate and is valid only at frequencies far below the self-resonant frequency of the coil (see Sec. 28). Moreover, at any one frequency we are at liberty to represent the coil either by a series combination of inductance and resistance or by a parallel combination (Fig. 47). The Qs given by the two representations are identical but the inductances are not. The two representations are related by the equations

$$R' = R(Q^2 + 1) \qquad (54a)$$

$$L' = L\left(\frac{Q^2 + 1}{Q^2}\right) \qquad (54b)$$

$$Q = \frac{\omega L}{R} = \frac{R'}{\omega L} \qquad (54c)$$

If $Q^2 \gg 1$ the factor

$$\frac{Q^2 + 1}{Q^2} \doteq 1$$

and

$$L' \doteq L$$

When $Q = 10$, L and L' differ by only 1 per cent and when $Q = 33$ they differ by only 0.1 per cent.

The *Hay bridge* (Fig. 47*c*), which differs from the Maxwell bridge only in that the capacitance-resistance combination in arm *b* is in series instead of in shunt, is also used to measure the inductance and Q of coils. In contrast to the Maxwell bridge, which is usually calibrated to read the inductance of a series RL representation of the coil, the Hay bridge can be conveniently calibrated to read the inductance of a paral-

lel RL representation. The balance equations are

$$R_d' = \frac{R_a R_c}{R_b} \tag{55a}$$

$$L_d' = R_a R_c C_b \tag{55b}$$

$$Q_d = \frac{R_d'}{\omega L'} = \frac{1}{\omega C_b R_b} \tag{55c}$$

The bridge dials can be calibrated to read L' and Q directly in a manner similar to that done with the Maxwell bridge.

38. Bridge Sensitivity and Equivalent Input and Output Circuits. The accuracy with which a bridge can be balanced or the sensitivity with which a small change in the impedance in one of the arms can be detected is determined by the power output of the oscillator, the sensitivity of the detector, and the rate of change of unbalance voltage with change in impedance. This sensitivity can be computed from the approximate equivalent circuits of the bridge valid for the case where the bridge is near balance. In the bridge circuit of Fig. 48*a* suppose that the impedances Z_a, Z_b, etc., are the values in the arms under balanced conditions and that ΔZ_d represents a small change in Z_d from the balanced condition.

When the bridge is balanced the potential difference between terminals BB' is zero so that this pair of terminals may either be short-circuited or be left as an open circuit without any effect on the input impedance between terminals AA'. The equivalent input impedance is therefore either

$$Z_{\text{input}} = \frac{(Z_a + Z_b)(Z_c + Z_d)}{Z_a + Z_b + Z_c + Z_d} \tag{56a}$$

or

$$Z_{\text{input}} = \frac{Z_a Z_d}{Z_a + Z_d} + \frac{Z_c Z_b}{Z_c + Z_b} \tag{56b}$$

The two forms are, of course, equivalent under balanced conditions and with Eq. (39) both reduce to

$$Z_{\text{input}} = \frac{Z_d(Z_a + Z_b)}{Z_a + Z_d} \tag{56c}$$

which is sometimes more convenient to use. In fact, since the four impedances of a balanced bridge are not independent but must satisfy the conditions of Eq. (39) any one of the four can be eliminated from Eq. (56*a*) or Eq. (56*b*) to give other expressions of the same form as Eq. (56*c*).

Under balanced conditions the potential between terminals AA' caused by a generator across terminals BB' must be zero from the reciprocity theorem. Therefore, the output impedance of a bridge can be written either as

$$Z_{\text{output}} = \frac{(Z_b + Z_c)(Z_a + Z_d)}{Z_a + Z_b + Z_c + Z_d} \tag{57a}$$

or

$$Z_{\text{output}} = \frac{Z_a Z_b}{Z_a + Z_b} + \frac{Z_c Z_d}{Z_c + Z_d} \tag{57b}$$

or

$$Z_{\text{output}} = \frac{Z_b(Z_a + Z_d)}{Z_a + Z_b} \tag{57c}$$

The open-circuited voltage between terminals BB' is the potential difference between these points due to the unbalance in the bridge and is

$$V_{0c} = V' \frac{Z_c Z_d}{(Z_c + Z_d)^2} \frac{\Delta Z_d}{Z_d} \tag{58a}$$

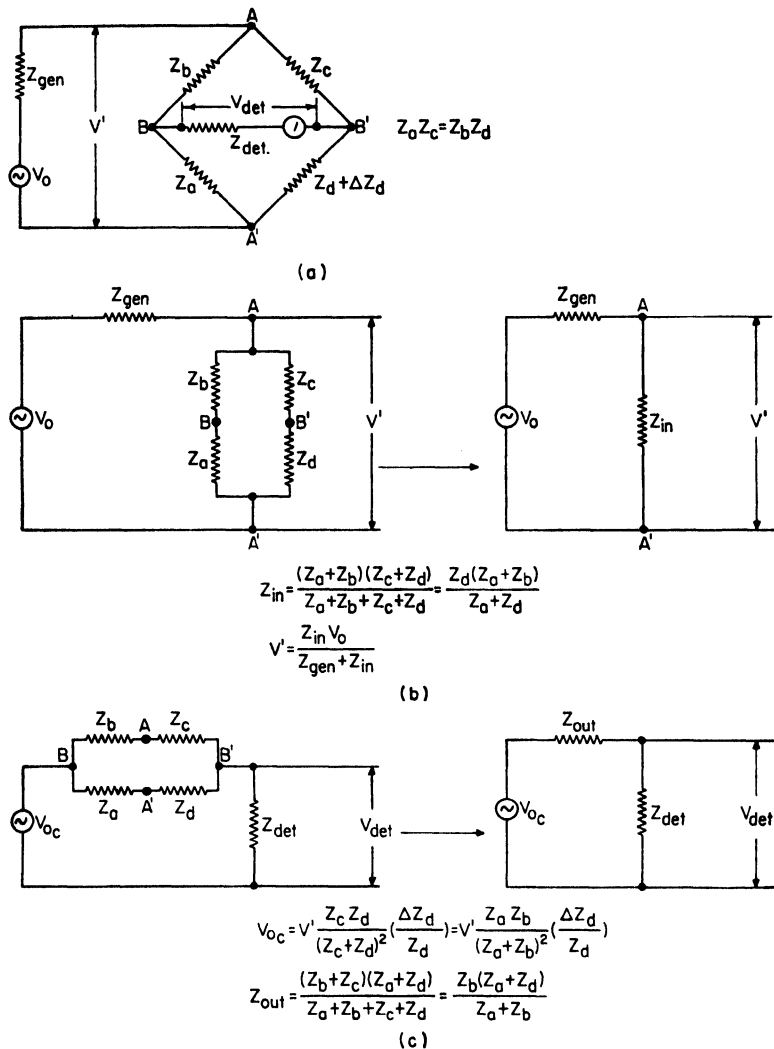



FIG. 48. Schematic diagram of bridge with generator and detector. (*a*) Equivalent input and output circuits of a balanced Wheatstone bridge; (*b*) equivalent input circuit at balance; (*c*) equivalent output circuit near balance.

With the aid of Eq. (39) the open-circuit voltage can also be expressed in terms of the impedances in the other two arms of the bridge:

$$V_{0c} = V' \frac{Z_a Z_b}{(Z_a + Z_b)^2} \frac{\Delta Z_d}{Z_d} \tag{58b}$$

Usually only the magnitude of V_{0c} is of interest.

From Eqs. (57) and (58) it is seen that the Thévenin equivalent output circuit for the bridge is that shown in Fig. 48*c*.

The equivalent circuits can be used to compute the signal generator voltage and detector sensitivity required to give any desired accuracy $\left|\frac{\Delta Z_d}{Z_d}\right|$ in the determination of Z_d. Suppose, for example, it is desired to find the generator voltage required to give a bridge sensitivity $\left|\frac{\Delta Z_d}{Z_d}\right|$ of 0.1 per cent with a detector having 5 μv sensitivity. First use the output equivalent circuit (Fig. 48c) and compute V' setting $V_{\text{det}} = 5 \times 10^{-6}$ volt and $\left|\frac{\Delta Z_d}{Z_d}\right| = 0.001$. Next, by means of the equivalent input circuit (Fig. 48b), compute V_0 using the value of V' just computed. The required open-circuit generator voltage to give the desired bridge sensitivity is V_0.

A consideration of the equivalent circuits of a bridge indicates the desirability of matching both the generator and the detector impedances to the input and output impedances, respectively, to optimize the bridge sensitivity. Aside from matching, it can be seen from Eq. (58b) that the sensitivity of a resistance-ratio bridge is maximized when the resistances in the ratio arms are made equal. Equation (58) also shows that the per unit sensitivity to unbalance of the smaller component of impedance is very much poorer than that for the larger when the ratio of the two components is large. For example, the sensitivity to unbalance of the resistive component of a high-Q coil is the same as that for the reactive component. However, the resistance is only $1/Q$ times as large as the reactance; therefore

$$\frac{\Delta R}{R} \doteq Q \frac{\Delta X}{X} \tag{59}$$

For the same reason the inductive balance is much more sensitive, in per cent, than the Q balance when a high-Q coil is measured on a Maxwell or a Hay bridge (Sec. 37). The relation is

$$\frac{\Delta Q}{Q} \doteq Q \frac{\Delta L}{L} \tag{60}$$

BRIDGE TECHNIQUES

39. Signal Generators and Detectors. Any oscillator or signal generator which will operate at the desired frequency and which will furnish sufficient power to give the required bridge sensitivity can be used as a signal source for bridge measurement. If the detector is a broad-band instrument, such as a VTVM, or a telephone headset commonly used at audio frequencies, then the harmonic output of the bridge oscillator should be low. The requirement of low harmonic content in the signal from the bridge oscillator is imposed because the impedances in the bridge arms change with frequency and, therefore, a bridge will not, in general, balance at the same dial settings for more than one frequency. The requirement for good wave form is not important if a narrow-band detector, such as a communication receiver, is used. At radio frequencies the signal may be amplitude-modulated with an audio tone to facilitate detection.

Any instrument sensitive to small voltages may be used as a bridge detector. A telephone headset is often used at audio frequencies and a sensitive VTVM may be used at all frequencies. A communications receiver whose sensitivity is of the order of a microvolt makes an excellent detector at radio frequencies. When a receiver is used as a bridge detector the automatic-volume-control circuits should be disabled. Narrow-band high-gain r-f receivers designed for use as bridge detectors are commercially available.

The power output of the oscillator and sensitivity of the detector required to give any bridge sensitivity can be computed from the equivalent circuits of Sec. 38.

In addition to having sufficient power and sensitivity, the oscillator and detector must be well shielded from each other for the true balance point to be detected. Poor shielding results in two effects: (1) the null is partially masked so that the detector

output shows a minimum rather than a true null, and (2) since the phase of the signal from the bridge varies very rapidly with bridge setting in the vicinity of a null, the minimum detector output occurs when the signal from the bridge tends to cancel the leakage signal. At radio frequencies it is necessary to use shielded cables to connect the detector and the oscillator to the bridge.

40. Bridge Transformers. Most signal generators and bridge detectors have large stray capacitances from the terminals to ground and in many cases one of the terminals is at ground potential. These stray capacitances appear across the bridge arms and at

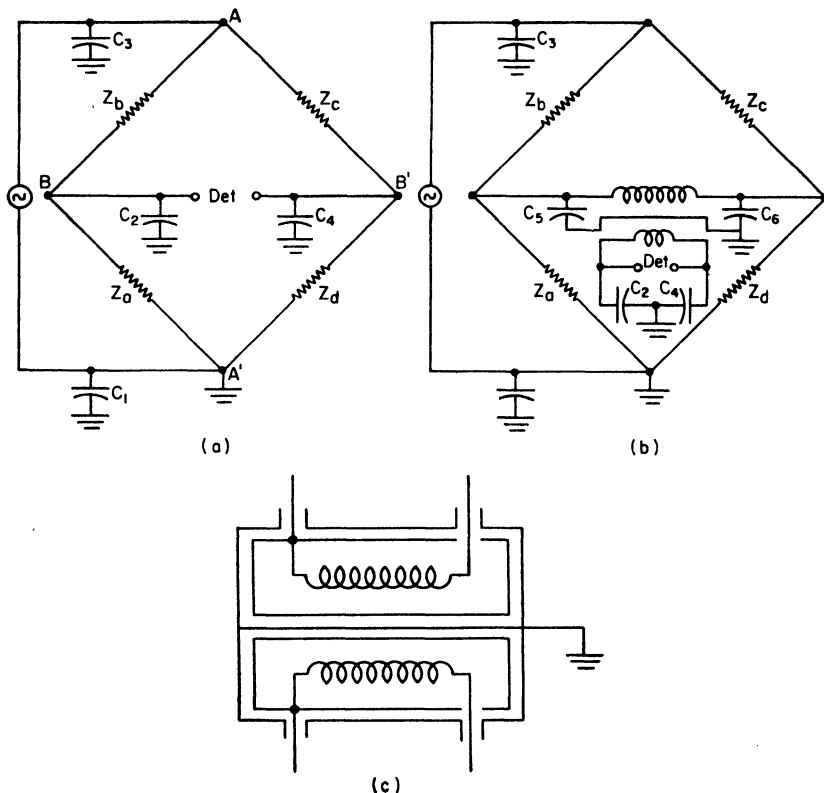



FIG. 49. (*a*) Schematic diagram showing effect of terminal capacitances in generator and detector; (*b*) bridge with shielded transformer; (*c*) typical transformer.

radio frequencies lead to erroneous results; even at audio frequencies they may cause serious errors. The use of both a detector and a signal generator with one terminal of each connected to ground is actually equivalent to grounding two corners of the bridge. Some provision must therefore be made for isolating the bridge from one of these equipments and a shielded transformer is usually used for this purpose. The transformer also permits the signal generator or detector to be matched to the input or output impedance of the bridge and thus increases the sensitivity.

The stray capacitances of the detector and oscillator in an unshielded bridge are illustrated in Fig. 49*a*, where C_1, C_2, C_3, and C_4 are the stray capacitances in question. Since point A' is grounded, C_1 has no effect. Likewise, C_3 is shunted across the generator terminals and has no effect on the bridge balance. On the other hand, C_2 and C_4 are shunted

across Z_a and Z_d, respectively, and enter into the balance equations of the bridge. The bridge balance is therefore dependent upon the stray capacitances in the detector circuit. Had point B been chosen for the ground point, then the stray capacitances in the generator would have entered into the balance equations.

The effect of the stray capacitances of the generator or detector can be removed by the use of a transformer with a shield between the windings (Fig. 49*b*). Since the capacitance between the primary and secondary winding can be reduced to a few tenths of 1 $\mu\mu$f the generator or detector capacitances to ground become negligible. The stray capacitances of the transformer secondary will now appear across the arms. The transformer can be designed, however, so that these stray capacitances are small and, moreover, they will remain fixed and can be included as part of the impedance in the respective arms. A typical shielded transformer is shown in Fig. 49*c*.

41. Bridge Errors. The lead inductances and shunt capacitances associated with the impedances of the bridge arms cause the most serious errors in bridge measurements. The effect of these stray reactances increases with frequency and is particularly important at radio frequencies. Even at audio frequencies serious errors may result if corrections are not made for these quantities. At radio frequencies bridge circuits must be designed to minimize the magnitudes of these stray reactances and provisions must be made for estimating their values so that corrections can be made in the bridge measurements.

Stray capacitances across the various bridge arms can be minimized to some extent by careful shielding. The most important function served by shielding, however, is to fix the magnitudes of these stray capacitances and to place them across points in the circuit where they can most easily be compensated or for which the errors introduced can most easily be corrected. Shielding, of course, includes the use of a shielded transformer as described in Sec. 40. An example of a shielded bridge is shown in Fig. 50.

Substitution of the unknown impedance in place of a standard in one arm of a balanced bridge provides one method for minimizing the effect of stray reactance. In this manner only the difference between the standard and the unknown need be measured. This method is nearly always used at high radio frequencies.

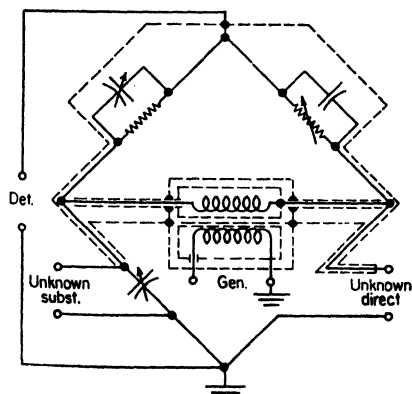



Fig. 50. General Radio shielded bridge.

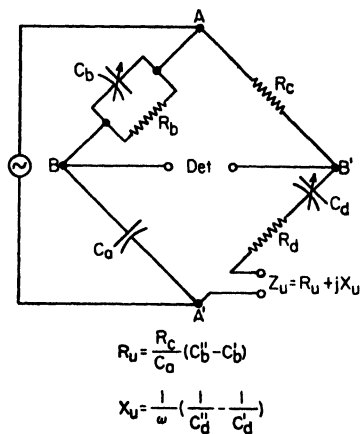



Fig. 51. Substitution method.

As an example of the substitution method, consider the Schering bridge of Fig. 51 where the standard impedance is a short circuit. The bridge is first balanced with a short circuit across the "unknown" terminals Z_u. The unknown impedance is inserted and the bridge is rebalanced. From the two sets of balance conditions the components of the unknown

impedance are found to be

$$R_u = \frac{R_c}{C_a}(C_b'' - C_b') \tag{61a}$$

and

$$X_u = \frac{1}{\omega}\left(\frac{1}{C_d''} - \frac{1}{C_d'}\right) \tag{61b}$$

where primes and double primes indicate the initial and final values of a parameter, respectively.

In this illustration the unknown impedance was inserted in series with the impedance in the bridge arm. It is also possible to place the unknown impedance in parallel with one of the arms, in which case the initial balance is usually made with the "unknown" terminals left as an open circuit.

42. Direct Capacitance. When a capacitor has terminal capacitances to a surrounding shield or to ground it can be represented as a three-terminal capacitor (Fig. 52*a*). The capacitance C_{12} is called the *direct capacitance* between terminals 1 and 2. The total capacitance between terminals 1 and 2 is the sum of the direct

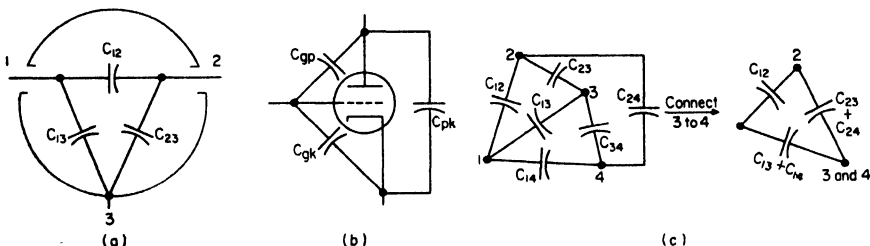



FIG. 52. (*a*) Shielded capacitor; (*b*) interelectrode capacitances of triode; (*c*) capacitance network and three-terminal equivalent network.

capacitance and the two capacitances C_{13} and C_{23} in series; and if the capacitor is connected to a capacitance bridge by terminals 1 and 2 with terminal 3 left disconnected the total capacitance will be the quantity measured. Often, however, the direct capacitance is the parameter of interest. The requirement for measuring direct capacitance in the presence of terminal capacitances may arise from any network which has more than two terminals. For example, any of the interelectrode capacitances of a triode tube (Fig. 52*b*) is the direct capacitance between two electrodes and must be measured in the presence of the other two interelectrode capacitances. Moreover, for the purpose of measuring direct capacitance, any network of capacitors can be reduced to a three-terminal network of the form shown in Fig. 52*a*, by connecting all the terminals together except the pair between which the direct capacitance is to be measured (Fig. 52*c*). The use of a guard circuit enables both direct capacitance and its dissipation factor to be measured.

43. Guard Circuit. When the four corners of a bridge are connected through impedances to a fifth point (Fig. 53*a*) the balance conditions will not be changed if the point P is kept at the same potential as the two corners B and B'. No potential difference exists across Z_2 and Z_4 in this case and they may be removed from the circuit leaving Z_1 in series with Z_3 across the generator terminals and the bridge balance unaffected. One set of conditions which gives a null on the detector is therefore

$$\frac{Z_a}{Z_b} = \frac{Z_d}{Z_c} = \frac{Z_1}{Z_3} \tag{62}$$

But there is a second set of conditions which gives a null on the detector. If the detector is connected across terminals AA' and the generator across BB' the detector

will show a null when

$$\frac{Z_a}{Z_d} = \frac{Z_b}{Z_c} = \frac{Z_2}{Z_4} \tag{63}$$

for the same reason given in obtaining Eq. (62). The reciprocity theorem, however, requires that the detector show a null if the generator and detector are interchanged. Therefore, the conditions of Eq. (63) also give a null when the detector is connected across terminals BB' and the generator across terminals AA'.

Both Eq. (62) and Eq. (63) contain the balance condition of the bridge without the guard circuit. Therefore, the balance of the bridge will be unaffected by the guard circuit if the guard circuit meets the requirements of either Eq. (62) or Eq. (63). Usually in using a guard circuit Z_1 and Z_3, which are across the generator, are balanced to satisfy Eq. (62), in which case these two impedances are often called the guard circuit while the impedances across the detector Z_2 and Z_4 are called *coupling impedances*.

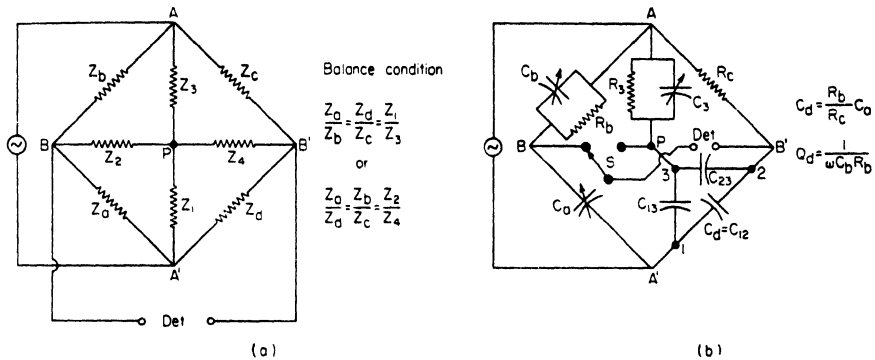



FIG. 53. (*a*) Schematic circuit; (*b*) Schering bridge and guard circuit for measuring direct capacitance.

When direct capacitance is measured by means of a guard circuit the two terminals between which the direct capacitance is to be measured are connected to the bridge in the usual manner so that the direct capacitance forms one arm of the bridge (Fig. 53*b*). The third terminal is then connected to point P.

The detector is first connected between points P and B' and the guard circuit is balanced. Then, by means of switch S, the detector is connected between points B and B' and the bridge is balanced. At balance the usual balance equations for the bridge apply, which for the Schering bridge of Fig. 53*b* are

$$C_d = \frac{R_b}{R_c} C_a \tag{64a}$$

$$Q_d = \frac{1}{\omega C_b R_b} \tag{64b}$$

In the particular case shown in Fig. 53*b* the guard-circuit balance is independent of the bridge balance since both Z_c and Z_d are held constant. Therefore, both the guard circuit and bridge need to be balanced only once, if the guard circuit is balanced first, to give a balance for both positions of the switch. If either Z_c or Z_d is changed in the process of balancing the bridge, as will occur when a substitution method is used, then the guard circuit and bridge balances are not independent and alternate readjustments of the guard circuit and bridge must be made until there is a balance for both positions of the switch.

44. Wagner Ground. As discussed in Sec. 40 the capacitances between ground and the terminals of the signal generator and the detector appear across some of the bridge arms and cause errors unless a shielded transformer is used. In making impedance measurements on a resistance-ratio bridge at audio frequencies satisfactory accuracy can often be obtained without the use of a transformer by means of a *Wagner ground* (Fig. 54). In using this circuit both the signal generator and the detector are left ungrounded. C_1 and C_3 are the terminal capacitances of the generator and C_2 and C_4 are those of the detector. These capacitances form a guard circuit which, if properly balanced, will leave the bridge balance unaffected. A low-impedance potentiometer with the movable tap grounded is connected across the generator terminals to provide a means of balancing the guard circuit. This potentiometer is in parallel with C_1 and C_3 between points A and A' and the movable tap is at the potential of point P. Since the impedance of the potentiometer is low compared with that of C_1 and C_3 the guard circuit is approximately balanced when

FIG. 54. Wagner ground.

$$\frac{R_1}{R_3} = \frac{R_a}{R_b} \tag{65}$$

The guard circuit and bridge are alternately adjusted with the switch S connected to ground and to point B', respectively, until a null exists for both positions of the switch. The usual balance equations for the bridge are then applicable.

A capacitive voltage divider can be used in place of the potentiometer shown in Fig. 54.

45. Mutual Inductance. In the measurement of mutual inductance a situation arises somewhat analogous to the measurement of direct capacitance (Sec. 42). In addition to the mutual inductance between the two windings of a transformer there are always associated with it the self-inductances of each of the windings. Although the usual transformer is a four-terminal network (Fig. 55*a*) one can connect terminals 3 and 4 together for the purpose of measurement. The equivalent circuit valid at frequencies where the shunt capacitances may be neglected is then as shown in Fig. 55*b*. As in the case of direct capacitance this is a three-terminal network; it is a T network, however, instead of a Δ network.

FIG. 55. Mutual inductance and equivalent circuit.

In measuring mutual inductance on a bridge the transformer is connected so that the T junction in the equivalent circuit forms one corner of the bridge with the mutual inductance forming part of one bridge arm and one of the other branches of the T forming part of an adjacent arm. The third branch of the T is then in series with either the detector or the generator where it does not enter into the balance equations.

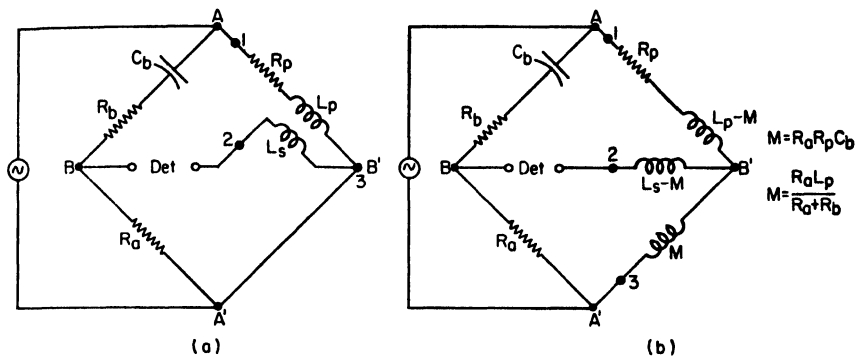



FIG. 56. Carey-Foster bridge. (a) Schematic diagram; (b) equivalent circuit.

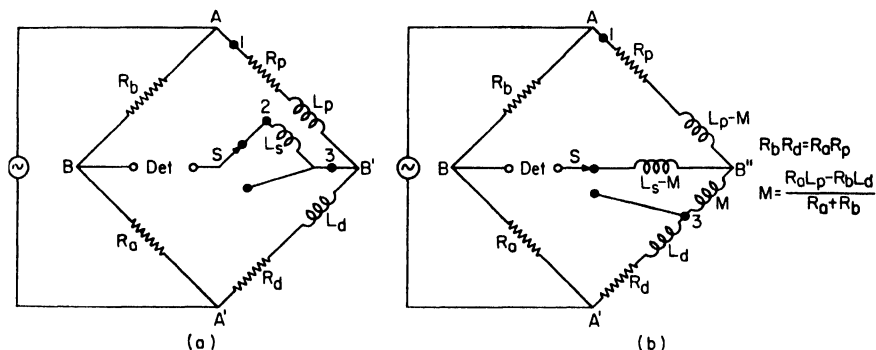



FIG. 57. Campbell bridge. (a) Schematic diagram; (b) equivalent circuit.

Two bridges commonly used for measuring mutual inductance are the *Carey-Foster bridge* (Fig. 56) and the *Campbell bridge* (Fig. 57). The balance equations for the Carey-Foster bridge are

$$M = R_a R_p C_b \tag{66a}$$

and

$$M = \frac{R_a L_p}{R_a + R_b} \tag{66b}$$

The mutual inductance enters into both the resistive and the reactive balances but is measured in terms of either the inductance or the resistance of the transformer primary, both of which must be measured on another bridge, by the Maxwell bridge of Fig. 46*b*, for example. In the Campbell bridge the self-inductance again enters into the determination of the mutual inductance. The balance equations for this bridge are

$$R_b R_d = R_a R_p \tag{67a}$$

and

$$M = \frac{R_a L_p - R_b L_d}{R_a + R_b} \tag{67b}$$

Changing the detector connection from terminal 2 to terminal 3 with switch S, however, permits L_p to be compared with L_d in a resistance-ratio bridge. The Campbell bridge, therefore, has the advantage that, although M is measured in terms of L_p, the primary inductance L_p can be determined by a second measurement on the same bridge. The fact that M is determined as the difference of two larger quantities, however, limits the accuracy of the measurements.

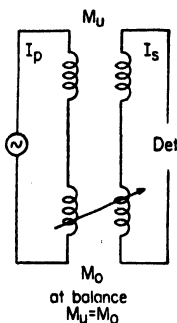



FIG. 58. Felici mutual-inductance balance.

An unknown mutual inductance can also be compared with a known calibrated variable mutual inductance on the *Felici mutual-inductance balance* (Fig. 58). In this circuit the two transformer primaries are connected in series to the signal source so that the same current flows through each primary. The detector is connected across the transformer secondaries, which are also connected in series. The secondaries, however, are connected so that the induced voltages are in opposition. The detector current is a null when the two induced voltages $j\omega M_0 I_p$ and $j\omega M_u I_p$ are equal. The balance condition is therefore

$$M_u = M_0 \tag{68}$$

46. Incremental Inductance. The inductance offered to the alternating component of current by an iron-cored coil which is used as a power-supply filter is a function of the direct current flowing through the coil. This inductance is called the *incremental inductance.*

In measuring incremental inductance it is necessary to send a direct current through the coil and to arrange the circuit so that this magnetizing current can be measured.

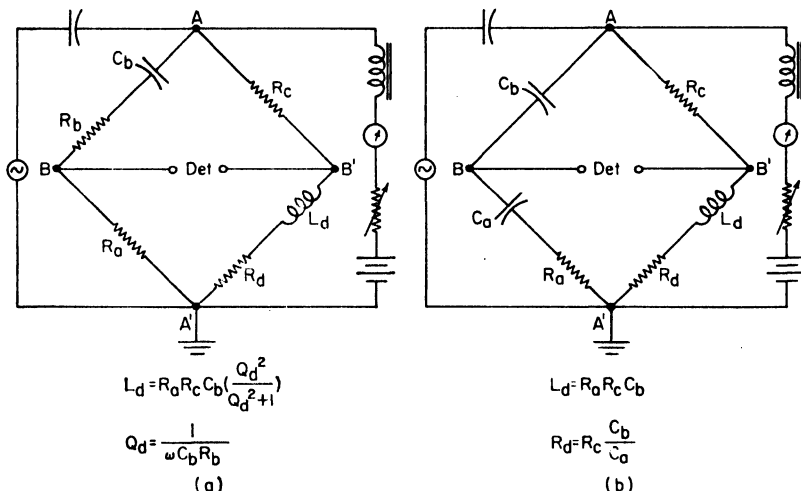



FIG. 59. Circuits for measuring incremental inductance. (*a*) Hay bridge; (*b*) Owen bridge.

The Hay bridge (Fig. 46*d*) or the Owen bridge (Fig. 46*e*) are usually chosen because the series capacitors block the direct current and permit it to be metered externally (Fig. 59).

T NETWORKS

In addition to the four-impedance bridge circuit two other null-type networks are sometimes used for impedance measurements. One of these is the *twin-T* or *parallel-T* network (Fig. 60*a*). The other is the bridged-T network (Fig. 60*b*). Both these cir-

cuits have the advantage over the bridge circuit that the generator and the detector have a common terminal which can be grounded. Hence a shielded transformer is not necessary.

47. Twin-T Network. When a null exists at the detector there is no potential difference between points B and G of the network shown in Fig. 60a. Therefore, the detector terminals may be short-circuited or left as an open circuit without affecting the condition of balance. In finding the balance condition it is convenient to assume

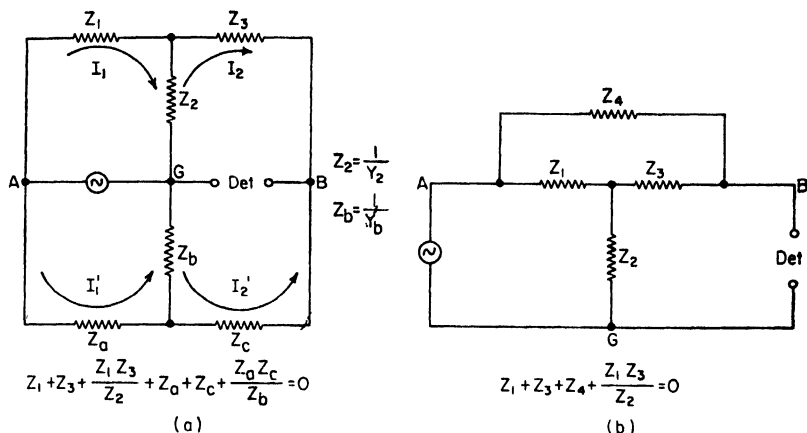



FIG. 60. T networks. (a) Twin-T circuit; (b) bridged-T circuit.

that a short circuit is across these terminals. The two Ts are then isolated and the output currents I_2 and I_2' can be found separately:

$$I_2 = \frac{VZ_2}{Z_1Z_2 + Z_2Z_3 + Z_1Z_3} \tag{69a}$$

$$I_2' = \frac{VZ_b}{Z_aZ_b + Z_bZ_c + Z_aZ_c} \tag{69b}$$

The null occurs when no current flows through the short circuit:

$$I_2 + I_2' = 0 \tag{70}$$

The balance condition is therefore

$$Z_1 + Z_3 + \frac{Z_1Z_3}{Z_2} + Z_a + Z_c + \frac{Z_aZ_c}{Z_b} = 0 \tag{71}$$

Since the impedances of which the network is composed are complex quantities, Eq. (71) actually represents a pair of equations in real variables. Moreover, it should be noted that although all the impedances contain positive real parts the product terms Z_1Z_3/Z_2 and Z_aZ_c/Z_b can have negative real parts so that the condition corresponding to the real part of Eq. (71) can be realized. The condition corresponding to the imaginary part of Eq. (71) can, of course, be realized since the component impedances can have either sign of reactance. In contrast to the bridge circuit the twin-T network cannot be used for d-c resistance measurements since a network composed only of resistances will not satisfy the conditions of balance expressed in Eq. (71).

48. Bridged-T Network. The bridged-T circuit (Fig. 60b) is a degenerate form of the twin-T network in which Z_b is an open circuit. The balance condition is therefore

given by Eq. (71) if we let $Z_b \to \infty$ and write Z_4 for the series combination $Z_a + Z_c$. The balance condition is

$$Z_1 + Z_3 + Z_4 + \frac{Z_1 Z_3}{Z_2} = 0 \tag{72}$$

49. Commonly Used T Networks. The twin-T circuit shown in Fig. 61 is very often used for measuring the admittance of high-impedance circuits at radio frequencies. This circuit gives good results for frequencies as high as 50 Mc and is more accurate than an r-f bridge when the impedance $1/Y_u$ is high. The balance equations for this circuit are

$$G_2 = \omega^2 C_1 C_3 \left(1 + \frac{C_b}{C_a}\right) R_c \tag{73a}$$

and

$$B_2 = -\omega C_1 C_3 \left(\frac{1}{C_1} + \frac{1}{C_3} + \frac{1}{C_a}\right) \tag{73b}$$

where the admittance between point A and ground is

$$Y_2 = \frac{1}{Z_2} = G_2 + jB_2 \tag{74}$$

As indicated in Fig. 61 a substitution method is nearly always used with this circuit. In this case the initial balance is made with the terminals u and u' left as an open circuit, after which the admittance to be measured Y_u is inserted between terminals u and u' and the circuit is rebalanced. The unknown admittance can be found from the relations

$$G_u = \frac{\omega^2 C_1 C_3 R_c}{C_a} (C_b'' - C_b') \tag{75a}$$

$$B_u = \omega(C_2' - C_2'') \tag{75b}$$

Fig. 61. Twin-T circuit for admittance measurements of high-impedance circuits at radio frequencies.

where the primes indicate values of a parameter at the initial balance and the double primes indicate the value of the same parameter at the final balance. In this arrangement only the circuit elements C_b and C_2 need to be made variable. C_b can be calibrated directly in conductance at any single frequency. The calibrated scale will then vary as the square of the frequency; it is thus seen that the range of conductances which can be measured decreases rapidly as the frequency is decreased. C_2 can be calibrated directly in susceptance at one frequency. The calibration then varies directly as the frequency.

It should be noted that the circuit of Fig. 61 has the advantage that one terminal of each of the variable circuit elements in the bridge and one terminal of the circuit element to be measured as well as one terminal of the generator and one terminal of the detector can all be grounded.

The twin-T network shown in Fig. 62 is equivalent to the Wien bridge (Sec. 37) and has similar balance equations:

$$\omega^2 = \frac{1}{R_b C_a C_c (R_1 + R_3)} \tag{76a}$$

$$\omega^2 = \frac{C_a + C_c}{R_1 R_3 C_2 C_a C_c} \tag{76b}$$

Equating the two expressions for ω^2 obtained from the balance condition gives an expression relating the capacitors and resistors in the network:

$$\frac{C_a + C_c}{C_2} = \frac{1}{R_b}\left(\frac{R_1R_3}{R_1 + R_3}\right) \tag{77}$$

When both the T networks are made symmetrical and when in addition C_2 is made twice C_a and R_1 is made twice R_b then both Eq. (75*a*) and Eq. (75*b*) reduce to the single expression

$$f = \frac{1}{2\pi R_1 C_a} \tag{78}$$

which is identical in form with Eq. (45) for the Wien bridge.

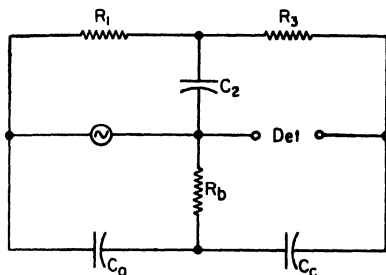



FIG. 62. Twin-T network equivalent to Wien bridge.

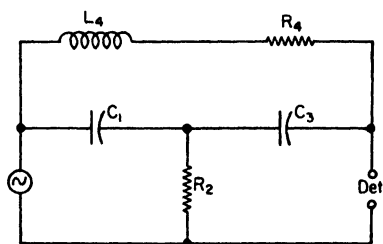



FIG. 63. Bridged-T circuit for measuring inductance.

The bridged-T circuit of Fig. 63 is useful for measuring inductance in terms of capacitance and resistance. For the measurement of the incremental inductance (Sec. 46) of iron-cored coils the d-c magnetizing current can be conveniently introduced in series with the a-c voltage, provided a low-resistance path is placed across the detector. The balance equations for this circuit are

$$R_4 = \frac{1}{\omega^2 C_1 C_3 R_2} \tag{79a}$$

$$L_4 = \frac{1}{\omega^2}\left(\frac{1}{C_1} + \frac{1}{C_3}\right) \tag{79b}$$

and

$$Q_4 = \frac{\omega L_4}{R_4} = \omega(C_1 + C_3)R_2 \tag{80}$$

In making measurements with this circuit it is convenient to make C_3 equal to C_1 and to gang them together. The balance equations [Eqs. (79*a*) and (79*b*)] then reduce to

$$R_4 = \frac{1}{\omega^2 C_1^2 R_2} \tag{81a}$$

$$L_4 = \frac{2}{\omega^2 C_1} \tag{81b}$$

and

$$Q_4 = 2\omega C_1 R_2 \tag{82}$$

IMPEDANCE MEASUREMENTS AT V.H.F.

Radio-frequency bridges with lumped circuit elements are useful for impedance measurements up to 150 Mc but above this frequency stray capacitances and inductances cause serious errors. On the other hand, slotted transmission lines become cumbersome below 300 Mc. Two null-type measuring instruments, which are not true bridges, are particularly useful for making impedance measurements in the intervening frequency range. One of these is the General Radio *admittance meter* and the other is the *Byrne bridge*.

50. General Radio Admittance Meter. The General Radio admittance meter is shown schematically in Fig. 64, where the coaxial lines normally used are drawn as two-wire lines. It consists of a transmission-line junction with four arms. Arms A, B,

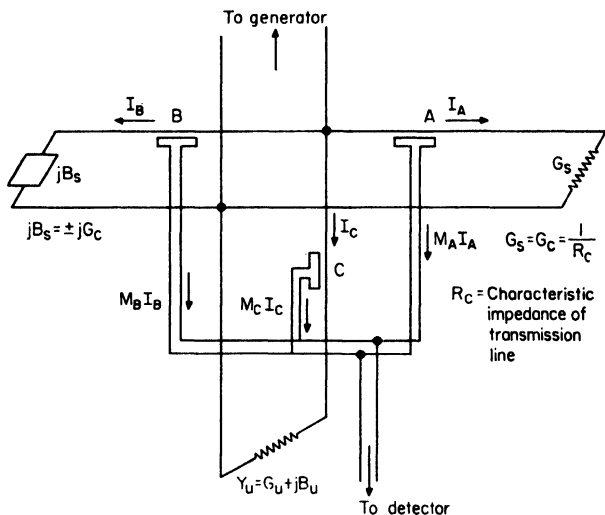



FIG. 64. General Radio admittance meter.

and C are identical and are terminated in a standard conductance, a standard susceptance, and the unknown admittance, respectively. An r-f source at the desired frequency is connected to the fourth arm. A small loop is located in each of the three symmetrical arms at points very near to, and equidistant from, the junction.

These three loops are connected to a common transmission line and thence to a detector. The current induced in each of these loops is proportional to the currents in the respective arm. The loops in the arms A and B can be rotated to vary the coupling between the loop and the line.

The principle of operation is as follows:

The same r-f voltage V is impressed on each of the arms A, B, C. The resulting currents in the three arms are

$$I_A = VG_s \tag{83a}$$

$$I_B = jVG_s \tag{83b}$$

$$I_C = VY_u = V(G_u + jB_u) \tag{83c}$$

and the current flowing into the detector junction is

$$I_d = V[M_AG_s + jM_BB_s + M_C(G_u + jB_u)] \tag{84}$$

where M_A, M_B, and M_C are the coupling ratios of the three loops. When the output of the detector is zero

$$G_u = \frac{-M_A}{M_C} G_s \tag{85a}$$

and

$$B_u = \frac{-M_B}{M_C} B_s \tag{85b}$$

Note that the measured admittance Y_u is the ratio of current to voltage at the position of the loop (point C). Likewise the standard conductance and susceptance are adjusted so that the ratios of I to V at the loop positions have the standard values.

51. Byrne Bridge. The basic circuit of the Byrne bridge is shown in Fig. 65. The signal generator is connected through a junction which contains two probes. One of these is a capacitive probe which samples the voltage and the other is a small loop whose output is proportional to the current. The coupling between each of these probes and the line can be varied by adjusting the distance between the probe and the center conductor. They are ganged together so that one probe moves toward the center conductor as the other moves away from it. The signals from these two probes are connected to opposite ends of the section of slotted transmission line which is terminated on both ends in its characteristic impedance to make reflections negligible. A movable probe on this line is connected to the detector. The detected signal is thus proportional to the sum of the two waves initiated by the voltage and current probes, respectively.

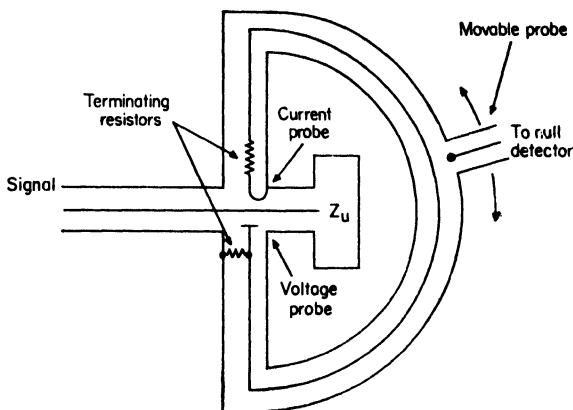



FIG. 65. Byrne bridge.

When the unknown impedance is a pure resistance the voltage and current at the junction are in phase with each other. However, the current loop is connected in such a manner that the two voltages applied to the opposite ends of the auxiliary section of the line are 180 deg out of phase with each other and a voltage minimum therefore occurs at the mid-point. When the probe coupling is adjusted so that the two voltages are of equal magnitude this minimum becomes a null. The ratio of the magnitudes of the current and voltage, which can be read from the calibrated probes, is then equal to the unknown resistance.

When the unknown impedance is complex the minimum is shifted from the center of the auxiliary line. The distance which the minimum is shifted is proportional to the difference in phase of the voltage and current at the junction and therefore to the phase angle of the impedance. This probe is calibrated to read phase angle directly at one frequency, but since the distance the minimum is shifted is also proportional to the wavelength, a correction factor must be applied at other frequencies. The ratio of the amplitudes of the voltage and current as read from the calibrated probe when the detector output is a null is the magnitude of the unknown impedance.

TRANSMISSION-LINE TECHNIQUES

Under steady-state conditions the voltage on a transmission line can be considered as the sum of two traveling waves, one traveling toward the termination, and the

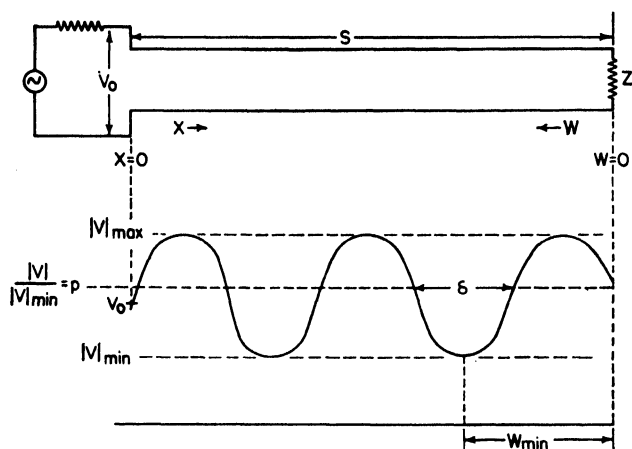



FIG. 66. Voltage standing-wave pattern on transmission line.

Table 1. Definitions of Symbols

Symbol	Definition
c	shunt capacitance per unit length
e	base of natural logarithms
g	shunt conductance per unit length
l	series inductance per unit length
n	an integer
p	a ratio (Fig. 66)
r	series resistance per unit length
R_c	characteristic resistance
r_1	real part of normalized impedance
s	distance from generator to termination (Fig. 66)
S	voltage-reflection coefficient
t	time
V_i	incident voltage wave (traveling toward termination)
V_r	reflected voltage wave (traveling toward generator)
V_0	voltage at generator end of line
w, w_{min}, w_{SC1}, w_{SC2}, w_{SC}, w', Δw	distance from termination (Figs. 66 and 72)
x	distance from generator (Fig. 66)
x_1	imaginary part of normalized function
y	shunt admittance per unit length
z	series impedance per unit length
Z_c	characteristic impedance
Z_1	normalized impedance of termination
α	attenuation constant
β	phase constant
γ	propagation constant
Γ	voltage-reflection coefficient of termination
δ	width across minimum of voltage distribution (Fig. 66)
ϵ	relative dielectric constant
θ	complex terminal function
λ	wavelength on line
ρ	attenuation function
Φ	phase function
ϕ_c	distortion factor
ψ	phase angle of reflection coefficient
ω	angular frequency

other a reflected wave traveling toward the generator. These two waves mutually reinforce at some points on the line and suffer destructive interference at others. The standing-wave pattern thus formed is determined entirely by the terminating impedance (Fig. 66) at the load end and can be used to find the magnitude and phase angle of this impedance. A section of transmission line properly instrumented so that the voltage distribution along it can be measured forms a very important impedance-measuring instrument at frequencies above 300 Mc. Below this frequency the equipment is cumbersome since a section of line approximately $\pi/2$ long is needed.

Table 2. Important Transmission-line Equations

A. Transmission-line parameters

$$Z_c = \sqrt{\frac{z}{y}}$$ is the characteristic impedance where z and y are the series impedance per unit length and the shunt admittance per unit length, respectively

$$Z_c = R_c(1 - j\phi_c)$$

$$Z_c = R_c\left(1 - j\frac{\alpha}{\beta}\right)$$ for lines with no dielectric losses

Usually $\phi_c \ll 1$ and

$$Z_c \doteq R_c$$

R_c is the characteristic resistance.
For a coaxial line:

$$R_c = \frac{138}{\sqrt{\epsilon}} \log_{10} \frac{b}{a}$$

where a = radius of inner conductor
b = radius of outer conductor
ϵ = relative dielectric constant of dielectric in space between a and b

For two-wire line:

$$R_c \doteq \frac{276}{\sqrt{\epsilon}} \log_{10} \frac{b}{a}$$

where a = radius of wires
b = distance between centers

B. Equations relating to the termination

$$z_1 = \frac{Z}{Z_c}$$ is the normalized impedance of the termination where Z is the impedance of the termination

$$z_1 = r_1 + jx_1$$
$$z_1 = \coth \theta$$

where $\theta = \rho + j\Phi$ = complex terminal function
ρ = attenuation function
Φ = phase function

$$r_1 = \frac{\sinh 2\rho}{\cosh 2\rho - \cos 2\Phi}$$

$$x_1 = \frac{-\sin 2\Phi}{\cosh 2\rho - \cos 2\Phi}$$

$$\rho = \frac{1}{2} \tanh^{-1} \frac{2r_1}{|z_1|^2 + 1}$$

$$\Phi = \frac{1}{2} \tan^{-1} \frac{-2x_1}{|z_1|^2 - 1}$$

$$\Gamma = \frac{z_1 - 1}{z_1 + 1}$$ = voltage-reflection coefficient

$$z_1 = \frac{1 + \Gamma}{1 - \Gamma}$$

$$\Gamma = |\Gamma| e^{j\psi}$$
$$\Gamma = e^{-2\theta}$$
$$|\Gamma| = e^{-2\rho}$$
$$\psi = -2\Phi$$

$$\rho = \frac{1}{2} \ln \frac{1}{|\Gamma|}$$

Table 2. Important Transmission-line Equations (*Continued*)

C. **Equations relating to the input impedance and reflection coefficient at any point along the line**

$$z_1(w) = \coth(\theta + \gamma w)$$

$$r_1(w) = \frac{\sinh 2(\rho + \alpha w)}{\cosh 2(\rho 2\alpha w) - \cos 2(\Phi + \beta w)}$$

$$x_1(w) = \frac{-\sin 2(\Phi + \beta w)}{\cosh 2(\rho + \alpha w) - \cos 2(\Phi + \beta w)}$$

$$\rho + \alpha w = \frac{1}{2}\tanh^{-1}\frac{2r_1(w)}{|z_1(w)|^2 + 1}$$

$$\Phi + \beta w = \frac{1}{2}\tan^{-1}\frac{-2x_1(w)}{|z_1(w)|^2 - 1}$$

$$\Gamma(w) = \frac{z_1(w) - 1}{z_1(w) + 1}$$

$$z_1(w) = \frac{1 + \Gamma(w)}{1 - \Gamma(w)}$$

$$S = \frac{1 + |\Gamma(w)|}{1 - |\Gamma(w)|}$$ is the voltage standing-wave ratio (VSWR)

$$S = \coth(\rho + \alpha w)$$

$$Z = Z_c^2 \frac{Z(w) - Z_{SC}(w)}{Z_c^2 - Z(w)Z_{SC}(w)}$$

where $Z_{SC}(w) = Z_c \tanh \gamma w$ is the impedance at point w with a short-circuit termination. This equation is convenient for computing Z when $Z(w)$ and $Z_{SC}(w)$ have been measured.

$$z_1 = \frac{z_1(w) - \tanh \gamma w}{1 - z_1(w) \tanh \gamma w}$$

For lossless lines:

$$z_1 = \frac{z_1(w) - j \tan \beta w}{1 - jz_1(w) \tan \beta w}$$

$$z_1(w) = \frac{1 + jz_1 \tan \beta w}{z_1 + j \tan \beta w}$$

and

$$z_1 = \frac{1 - jS \tan \beta w_{\min}}{S - j \tan \beta w_{\min}}$$

The last equation is sometimes convenient for computing the impedance of a termination when the VSWR and the position of a minimum are known.

52. Theory of Transmission-line Impedance Measurements.[16,17,18] Before discussing transmission-line techniques for measuring impedance, it is necessary to digress briefly to define several parameters which are needed to describe these techniques. Definitions of the symbols used are shown in Table 1 and the more important equations are given in Table 2.

The electrical characteristics of a two-conductor transmission line are most often described by a pair of complex quantities called the *characteristic impedance* and the *propagation constant.* The characteristic impedance Z_c is the impedance which would be measured at the input terminals to the line if the line extended on indefinitely. It is the ratio of voltage to current in a wave traveling down the line. When a transmission line is terminated in its characteristic impedance the traveling wave is completely absorbed at the termination and the line behaves as if it extends indefinitely. The propagation constant γ describes the propagation of a wave along the semi-infinite line. If the voltage at the input were $V_0e^{j\omega t}$ then the voltage at a point x meters down the line would be $V_0e^{(j\omega t - \gamma x)}$. Since $\gamma = \alpha + j\beta$, the voltage along the line is given by

$$V(x,t) = V_0e^{-\alpha x + j(\omega t - \beta x)} \tag{86}$$

This represents a voltage which is attenuated by a factor $e^{-\alpha x}$ and is retarded in phase by βx radians. The real part of γ is therefore called the *attenuation constant* and the imaginary part is called the *phase constant.*

The distance along the semi-infinite line in which the phase of the traveling wave undergoes a phase shift of 2π radians is called the wavelength. The wavelength λ is related to the propagation constant by

$$\beta\lambda = 2\pi \tag{87}$$

Alternatively, the characteristics of a transmission line can be described by another pair of complex quantities—the series impedance per unit length z and the shunt admittance per unit length y. These two pairs of complex quantities which are commonly used to characterize a two-conductor transmission line are related by the following equations:

$$Z_c = \sqrt{\frac{z}{y}} \tag{88a}$$

$$\gamma = \sqrt{zy} \tag{88b}$$

where

$$z = r + j\omega l \tag{89a}$$
$$y = g + j\omega c \tag{89b}$$

The series inductance per unit length is l and the shunt capacitance per unit length is c.

If $r \ll \omega l$ and $g \ll \omega c$, then Z_c is very nearly a pure resistance

$$Z_c \doteq R_c$$

and γ is almost a pure imaginary

$$\gamma \doteq j\beta$$

Conversely, if the attenuation along the line is very small ($\alpha \ll \beta$), then

$$Z_c \doteq R_c$$

When a transmission line is terminated in an arbitrary impedance Z_a, the wave traveling down the line is reflected at this termination and travels back along the line with the same attenuation and phase constants as before. The ratio of the reflected voltage wave to the incident wave at the termination is called the voltage-reflection coefficient of the termination and is designated by Γ. Since the wave may undergo phase change as well as suffer attenuation on reflection, the reflection coefficient is a complex quantity. It is related to the terminating impedance by

$$\Gamma = \frac{Z - Z_c}{Z + Z_c} \tag{90}$$

The reflection coefficient for a terminating impedance equal to Z_c is zero; for an open-circuit termination $\Gamma = 1$ and for a short circuit $\Gamma = -1$. Any termination consisting of a pure reactance has a reflection coefficient whose magnitude is unity if the attenuation constant of the transmission line is small ($\alpha \ll \beta$; therefore $Z_c = R_c$).

It is convenient to normalize the terminating impedance to the characteristic impedance of the line; thus

$$\frac{Z}{Z_c} = z_1 \tag{91}$$

In terms of the normalized impedance

$$\Gamma = \frac{z_1 - 1}{z_1 + 1} \tag{92}$$

When a voltage V_0 is impressed at the input terminals of a transmission line of length s terminated by an impedance with reflection coefficient Γ the voltage along this line is the sum of two waves—an incident wave

$$V_i(x,t) = V_0 e^{j\omega t - \gamma x} \tag{93}$$

traveling toward the termination and a reflected wave

$$V_r(x,t) = \Gamma V_0 e^{j\omega t - \gamma s} e^{-\gamma(s-x)} \tag{94}$$

traveling toward the generator. In using the voltage distribution on the line for determining the impedance of the termination it is convenient to express this voltage in terms of w, the distance from the termination to the point under investigation. Then

$$w = s - x$$

and the voltage on the line is

$$V(w,t) = V_0 e^{j\omega t - \gamma s}(e^{\gamma w} + \Gamma e^{-\gamma w}) \tag{95}$$

The ratio of reflected voltage to incident voltage at any point w corresponds to an apparent coefficient of reflection at the point w of

$$\Gamma(w) = \Gamma e^{-2\gamma w} \tag{96}$$

The apparent normalized impedance of the termination as seen from the point w is therefore

$$z_1(w) = \frac{1 + \Gamma(w)}{1 - \Gamma(w)} = \frac{1 + \Gamma e^{-2\gamma w}}{1 - \Gamma e^{-2\gamma w}} \tag{97}$$

Equations (95), (96), and (97) take on a much simpler form and computations involving their use can be carried out more conveniently if the voltage-reflection coefficient is expressed as a complex exponential:

$$\Gamma = e^{-2\theta} \tag{98}$$

where

$$\theta = \rho + j\Phi \tag{99}$$

The quantity θ is called the *complex terminal function*, its real part is called the *attenuation function*, and its imaginary part is called the *phase function*.

Using this notation Eq. (95) becomes

$$V(w,t) = 2V_0 e^{j\omega t - (\theta + \gamma s)} \cosh(\theta + \gamma w) \tag{100}$$

and Eq. (97) becomes

$$Z_1(w) = \coth(\theta + \gamma w) \tag{101}$$

The normalized impedance of the termination is

$$Z_1 = \coth \theta \tag{102}$$

Both the real and imaginary parts of the terminal function θ can be determined from the magnitude of the voltage distribution, and the normalized impedance can then be computed from Eq. (102). The phase function Φ can be found from the position of a minimum or a maximum on the measuring line. From Eq. (100) the magnitude of the voltage on the line is

$$|V|^2 = 4V_0^2 e^{-2(\rho + \alpha s)}[\cos^2(\Phi + \beta w) + \sinh^2(\rho + \alpha w)] \tag{103}$$

The attenuation on a good measuring line is very low and under this condition the minimums of $|V|^2$ occur at

$$\Phi = (2n + 1)\frac{\pi}{2} - \beta w_{\min} \tag{104}$$

and the maximums occur at

$$\Phi = n\pi - \beta w_{\min} \tag{105}$$

where n is an integer.

A measurement of the distance to a minimum or a maximum thus permits Φ to be determined. The position of a minimum is almost always used since it can be located much more accurately than can a maximum.

There are two common methods for determining the attenuation function. One is the *standing-wave-ratio method* and the other is the *curve-width method.* In the standing-wave-ratio method ρ is determined from the standing-wave ratio, which is defined as the ratio of a maximum to a minimum of the voltage distribution.

$$S = \frac{|V|_{\max}}{|V|_{\min}} \tag{106}$$

Substitution of the conditions for a maximum and minimum from Eqs. (104) and (105) into Eq. (103) and taking the square root of the ratio yield

$$S = \coth\,(\rho + \alpha w_{\min}) \tag{107}$$

where $w_{\min}$ is the distance from the termination to the position of the minimum and where the difference in attenuation between the position of the maximum and the position of the minimum is neglected. The quantity $\alpha w_{\min}$ is also usually negligible in comparison with ρ, but this quantity can be measured and taken into account if required. The attenuation function ρ can be found from S by means of Eq. (107).

In the curve-width method ρ is determined from the distance between two points on either side of a minimum where the magnitude of the voltage has risen to p times the minimum value. When the ratio

$$\frac{|V|^2}{|V|_{\min}^2} = p^2 \tag{108}$$

is formed from Eq. (103) the following result is obtained:

$$(p^2 - 1)\sinh^2\,(\rho + \alpha w_{\min}) = \sin^2\frac{\beta\delta}{2} \tag{109}$$

where δ is the distance from the point where $\dfrac{|V|}{|V|_{\min}} = p$ on the other side (Fig. 66). The ratio $p = 2$ is usually chosen, in which case

$$\sinh\,(\rho + \alpha w_{\min}) = \sin\frac{\beta\delta}{2} \tag{110}$$

For standing-wave ratios greater than 10 the hyperbolic and the trigonometric sines in Eq. (110) are approximately equal to their arguments and

$$\alpha w_{\min} \doteq \frac{\beta\delta}{2} \tag{111}$$

Under this restriction

$$S \doteq \frac{2}{\beta\delta} = \frac{\lambda}{\pi\delta} \tag{112}$$

The standing-wave-ratio method is almost always used when the standing-wave ratio is less than 10. The curve-width method is preferable for standing-wave ratios greater than 10.

53. Standing-wave-ratio and Impedance Measurements. The measuring line used for making standing-wave and impedance measurements consists of a section of transmission line with an axial slot through which a small probe projects. This probe is arranged so that it can be moved along the line while its depth of projection into the slot remains constant. A scale attached to the line permits the position of the probe

to be determined. A signal source is connected to one end of the line and the termination whose impedance is to be measured is connected to the other. The voltage distribution on the line is determined by observing the output of a detector connected to the probe as the probe is moved (Fig. 67).

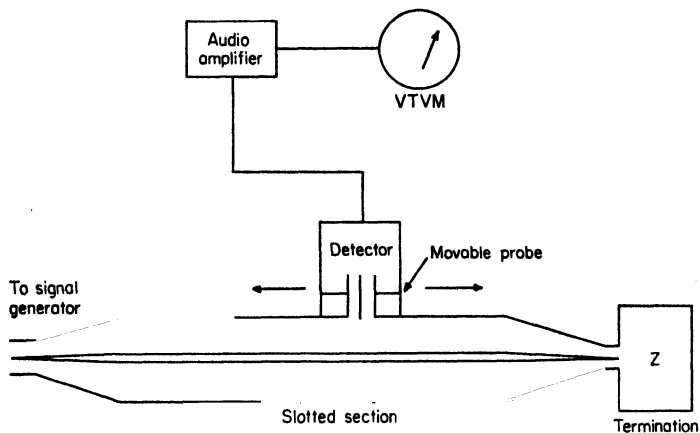



FIG. 67. Block diagram of measuring line and associated equipment.

The standing-wave ratio can be measured directly and the terminal functions ρ and Φ from which the impedance is computed can be easily determined from the voltage distribution.

a. Determination of the Phase Function Φ. The phase function is determined as follows:

1. Place a short circuit at the terminal end of the line.

2. Locate two adjacent minimums on the slotted section—preferably the first and second from the termination. The distance between these two minimums is $\lambda/2$:

$$\lambda = 2(w_{sc2} - w_{sc1}) \tag{113}$$

and each of the minimums is an integral number of half wavelengths from the termination. Since the scale may not be zero at the termination the position of one of the short-circuit minimums serves to locate the termination with respect to the scale.

3. Remove the short circuit and connect the unknown termination in its place.

4. Find the position of the new minimum nearest any one of the short-circuit minimums. The new minimum must be less than $\lambda/4$ from the reference short-circuit minimum. Then

$$\Phi = \frac{\pi}{2} + \frac{2\pi}{\lambda}(w_{sc} - w') \tag{114}$$

where w_{sc} = position of reference short-circuit minimum
w' = position of minimum with unknown termination

b. Standing-wave-ratio Method of Determining Attenuation Function ρ. The detectors used in making slotted-line measurements are usually crystals or bolometers and the detected output is proportional to the square of the voltage on the line. A linear detector or one following any other law may be used but the detection law must be known.

The procedure for determining the standing-wave ratio is as follows:

1. With the unknown termination connected to the slotted section, set the probe to the position of a minimum and adjust the detector and signal source power so that the detector output is well above the noise level.

2. Read the detector output.

3. Without changing the gain of the detector or signal generator output setting, move the probe to the position of a maximum and read the detector output. The maximums are

broad so that the exact position of a maximum is difficult to determine but the maximum value of the detector output as the probe is moved along the line can be determined easily.

4. The voltage standing-wave ratio

$$S = \frac{|V|_{max}}{|V|_{min}} \tag{106}$$

If a square-law detector is used

$$S = \sqrt{\frac{\text{maximum detector output (step 3)}}{\text{minimum detector output (step 2)}}}$$

If a linear detector is used

$$S = \frac{\text{maximum detector output (step 3)}}{\text{minimum detector output (step 3)}}$$

The terminal function ρ can be determined from the relation

$$S = \coth (\rho + \alpha w_{min}) \tag{107}$$

where w_{min} is the distance from the minimum used to the termination. The quantity αw_{min} can usually be neglected so that

$$S \doteq \coth \rho \tag{115}$$

The chief source of error in this method of determining ρ is the accuracy of the detector. In crystal detectors the detection law changes with signal level and becomes linear at large signal levels. It is therefore necessary to calibrate the detector if high standing-wave ratios are to be measured or if extreme accuracy is desired. Bolometer detectors follow the square law much more accurately than do crystals but are usually less sensitive and require more power from the signal source.

c. Curve-width Method of Determining the Terminal Function ρ

1. With the unknown termination connected to the slotted section set the probe to the position of a minimum and adjust the detector and signal-source power so that the detector output is well above the noise level.

2. Locate the two points on each side of the minimum where the voltage is 1.41 times the minimum value. When a square-law detector is used the detector output is double the minimum value at these points. Record the distance δ between these points.

3. Repeat steps 1 and 2 with a short circuit in place of the unknown termination and again record the distance δ'.

4. For standing-wave ratios greater than 10

$$\rho \doteq \frac{\pi(\delta - \delta')}{\lambda} \tag{116}$$

and

$$S \doteq \frac{1}{\rho} \tag{117}$$

This method is usually more accurate for standing-wave ratios greater than 10 than is the standing-wave-ratio method. However, the accuracy is dependent upon the accuracy with which the two points at which the voltage is 1.41 times the minimum value can be located. The accuracy can be improved if the voltage at 8 or 10 points in the vicinity of a minimum is measured and the voltage distribution is plotted on paper. The width between the 1.41 voltage points can then be determined from this plot.

54. The Measuring Line. The measuring line usually consists of a section of coaxial line with an axial slot through which a small movable probe projects (Fig. 68). Although this piece of equipment is basically very simple, the accuracy of the measurements made through its use depends to a large extent on the design of the line and the precision with which it is constructed.

In designing a measuring line, or in choosing one from those commercially available, the following factors should be considered:

1. The line should be of rigid mechanical construction and should have extremely low losses. Preferably it should have air dielectric and the slot should be narrow to prevent radiation losses.
2. The characteristic impedance should be constant along its length. This means that the cross section of the line should be constant along its length and the inner conductor should be accurately centered. The value of characteristic impedance chosen is arbitrary, and it is usually convenient to make this impedance equal to the characteristic impedance of the coaxial power-transmission cables with which the measuring line is usually used.
3. Mechanical tolerances permitted in maintaining constant characteristic impedance and in keeping the probe coupling constant are less severe on large lines. However, the diameter of the line must be small enough so that all modes except the *TEM* mode are suppressed at the highest frequency for which the line is to be used.
4. Supports for the center conductor should be designed so that reflections from them are negligible. Dielectric supports within the slotted section should be axially continuous along the length of the slot. A wedge-shaped support of this type is satisfactory (Fig. 68*b*), or the inner conductor can be left unsupported through the slotted region if it is sufficiently rigid, in which case the dielectric supports at the ends can be undercut to

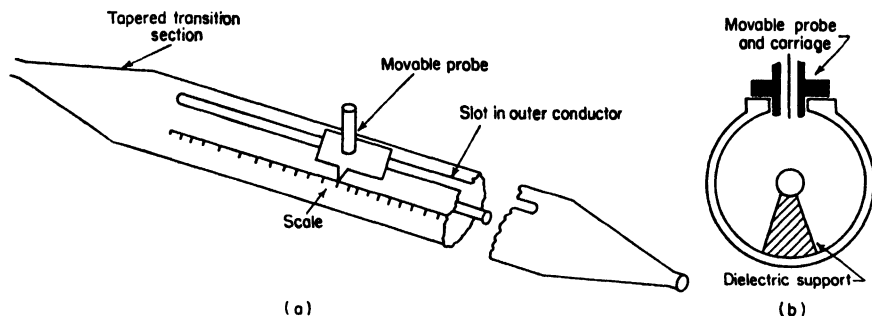



FIG. 68. Measuring line and cross-sectional view.

minimize reflections. Foam-polystyrene supports, which cause only slight reflections, are often used.

5. The tapered transition section and connector by which the measuring line is connected to coaxial cables normally used for power transmission at these frequencies should be designed to have negligibly small reflections.
6. The probe carriage should be designed so that the depth of penetration of the probe into the slot is kept constant as the probe is moved along the line. Changes in probe depth and the resultant changes in coupling to the line cause inaccuracies when low standing-wave ratios are measured and they limit the minimum VSWR which can be obtained. Small circumferential movements of the probe also change the probe coupling. Tolerances on circumferential motion are not so severe as those for radial motion and the errors caused by circumferential motion can be reduced by extending the coaxial shield around the probe through the wall of the outer conductor of the line (Fig. 68*b*).

The variation of the coupling of the probe to the line as the probe is moved axially can be measured by means of an audio-frequency signal. In this test an a-f generator is connected to the line with an open-circuit termination on the load end and the output of the probe (without detector) is monitored with a VTVM after amplification in a high-gain audio amplifier. The wavelength at audio frequencies is so long that the voltage distribution on the line is constant and any changes in the output of the amplifier as the probe is moved are due to changes in coupling.

An alternative configuration (Fig. 69) often used for a measuring line has a cylindrical inner conductor and an outer conductor consisting of two parallel flat plates. Although the slot is wide it is in a region of low electric field and radiation losses are low. This line has the advantage that the probe is in a region of low field and that the

probe penetration is large for the desired coupling. Small changes in probe depth therefore have negligible effect on the coupling. The design of transition sections for connecting the measuring line to coaxial cables poses a problem, but lines of this type are commercially available in which the reflections from the transition section are extremely low.

Fig. 69. Parallel-plate line.

55. Signal Generators and Detectors. Any oscillator or signal generator which will operate at the desired frequency and which will furnish sufficient power to give a signal about 10 db above the detector noise at the minimum of the voltage distribution can be used as a signal source for transmission-line measurements. The harmonic output of the signal source should be low and it is advisable to put a low-loss filter, with a cutoff frequency between the fundamental and second harmonic, in the output line to ensure low harmonic content, particularly if high standing-wave ratios are to be measured. It is usually convenient to use an r-f signal modulated with an audio tone.

The most common type of detector consists of a crystal or a bolometer to detect the audio modulation followed by a high-gain narrow-band audio amplifier. Some tuning device such as a double-stub tuner is usually necessary to obtain sufficient sensitivity. Often the bolometer or crystal holder and associated tuning device are built into the probe carriage. The output from the audio amplifier can be measured with a VTVM. Audio amplifiers with built-in voltmeters designed specifically for transmission-line measurements are commercially available.

A microammeter can be used with a crystal detector to measure the rectified crystal current and thus indicate the voltage on the line. This system is not very sensitive and requires more power from the signal source than does the crystal and audio amplifier. A superheterodyne receiver makes an excellent detector but the high sensitivity of this instrument is not usually needed and the comparatively simple crystal or bolometer with audio amplifier is more convenient to use.

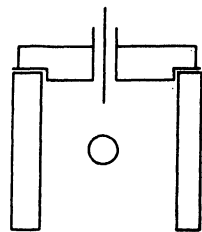
The detection law must be known to measure the VSWR. In this respect bolometers are superior to crystals. The detected output from a bolometer is usually proportional to the square of the input signal over a wide range of input signal levels ($\sim$30:1). A crystal is only approximately a square-law detector and the detected output can be assumed proportional to the square of the input signal with reasonable accuracy only over an input range of 3 or 4 to 1. A crystal, however, is usually more sensitive than a bolometer.

Fig. 70. Use of voltage distribution for detector calibration.

The detector can be calibrated by means of the current distribution on the slotted section of transmission line when the line is terminated in a short circuit. The distribution of current on a short-circuited line is sinusoidal with high accuracy except in the regions very near the minimums.

The procedure for calibration is as follows (Fig. 70):

1. Locate two adjacent minimums on the line. The distance between these minimums is $\lambda/2$.
2. Find the point A halfway between these two minimums; all distances x are to be measured from this point.
3. Set the probe at point A and adjust the input signal level (signal source) until the output from the detector is the maximum needed for measurements. Designate the voltage at point A by V.
4. Move the probe (in either direction) in equal increments of cos $(2\pi x/\lambda)$ until

$$\cos \frac{2\pi x}{\lambda} = \tfrac{1}{2}$$

($x/\lambda = 1/6$) and record the output voltage for each probe setting. The input voltages are given by

$$V_i = V \cos \frac{2\pi x}{\lambda}$$

5. Set the probe at point A again and adjust the input signal until the output is the same as it was with the probe set at $x/\lambda = 1/6$ in step 4. The input voltage at point A is now $V/2$ and the voltage distribution is $V_i = V/2 \cos (2\pi x/\lambda)$.

6. Repeat steps 4 and 5 until the signal decreases to the detector noise level, or as far as is desired.

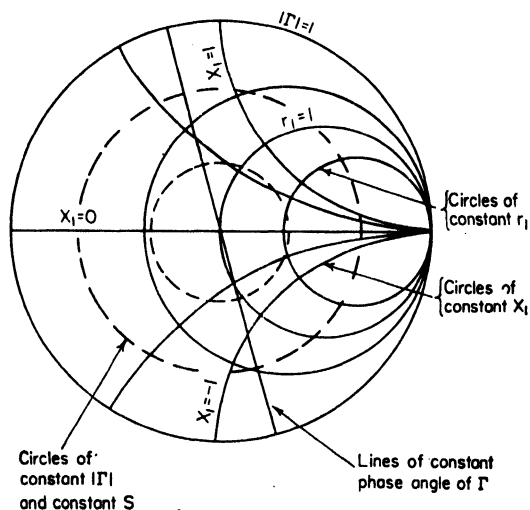
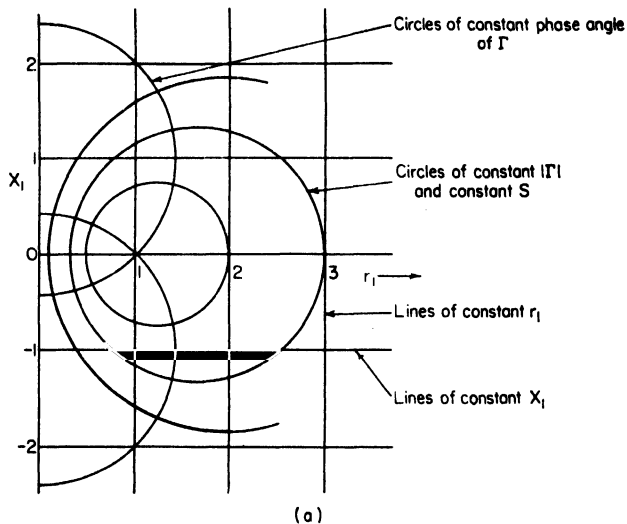



FIG. 71. Transmission-line charts. (a) r_1-X_1 chart; (b) Smith chart.

56. Transmission-line Charts.[18] Several types of diagram are in common use for graphical solution of transmission-line problems. Graphical solution is much faster than numerical solution of the equations, and although it does not afford as high accuracy, the accuracy is sufficient for most computations.

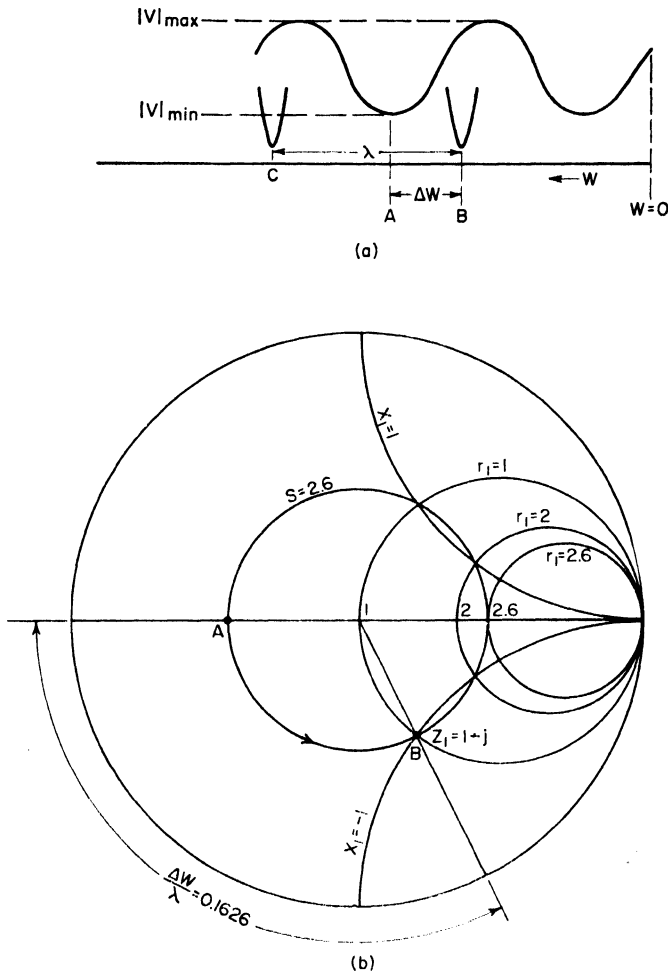



FIG. 72. Use of the Smith chart. (*a*) Voltage distribution; (*b*) graphical computation of impedance.

One of the simplest charts is that shown in Fig. 71*a*. It consists of a rectangular coordinate system in which one coordinate is the real part and the other the imaginary part of the normalized impedance $z_1 = Z/Z_c$. Only the right half plane is needed since negative values or r_1 are not encountered (assuming Z_c real). Superimposed on this coordinate system are lines of constant magnitude of the reflection coefficient and lines of constant phase angle of the reflection coefficient. From Eqs. (96) and (98)

the apparent reflection coefficient at any point on a line is

$$\Gamma_{(w)} = \Gamma e^{-2\alpha w} e^{-2j\beta w} \tag{118a}$$
$$\Gamma_{(w)} = e^{-2(\rho+\alpha w)} e^{-2j(\Phi+\beta w)} \tag{118b}$$

where Γ is the reflection coefficient at the load. Therefore, moving along a line of constant $|\Gamma|$ on the chart corresponds to moving along a lossless transmission line and the distance moved determines the change in the angle βw. The lines of constant $|\Gamma|$ are also lines of constant VSWR,

$$S = \frac{1 + |\Gamma|}{1 - |\Gamma|} \tag{119}$$

and are usually marked directly in S.

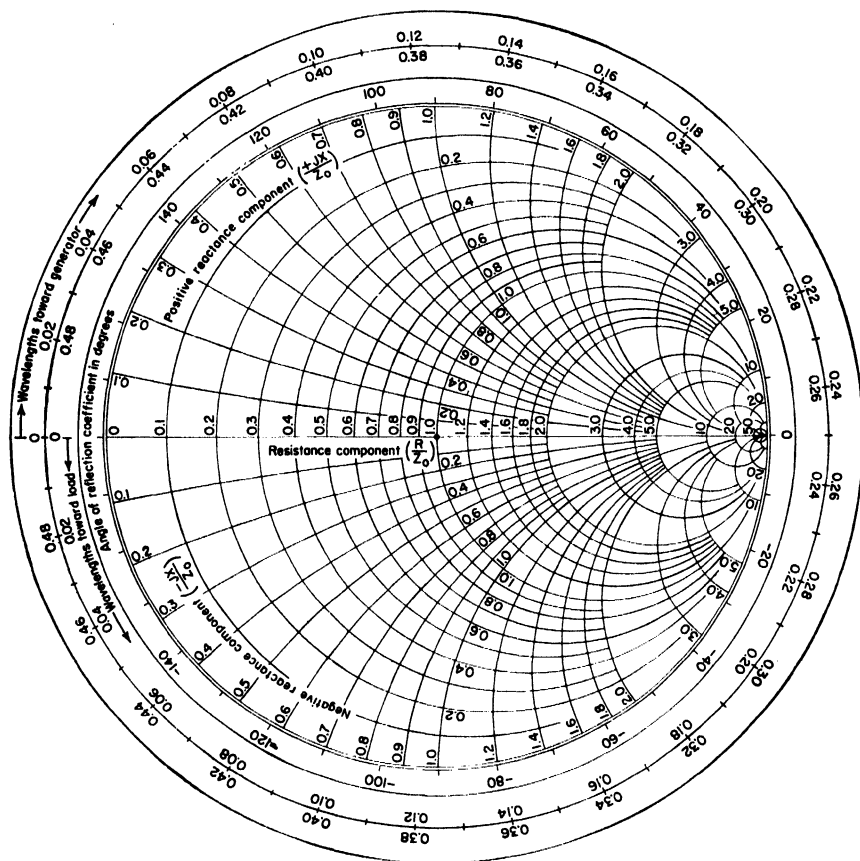



FIG. 73. Smith chart.

The Smith chart[19] (Fig. 71*b*) is the most widely used transmission-line chart and is probably the most convenient diagram to use. It is, in effect, the chart of Fig. 71*a* rolled up so that the whole right half plane of the r_1-x_1 coordinate system is contained in a unit circle. The lines of constant r_1 and lines of constant x_1 are plotted on a polar coordinate system in which the radial coordinate is $|\Gamma|$ and the angular coordinate is the phase angle of Γ.

Transforming an impedance through a section of lossless transmission line, therefore, consists of moving around the chart at constant radius. Lines of constant VSWR are circles of constant radius (usually not drawn on the diagram) and the value of r_1 at which one of these circles crosses the positive real axis is the value of S for the circle. The value of r_1 at which it crosses the negative real Γ axis is $1/S$. The angular coordinate is $2(\Phi + \beta w)$ as can be seen from Eq. (118b) and it is usually marked directly in w/λ so that the length of line (in wavelengths) through which a transformation is made can be read directly from the chart.

As an example of the use of the Smith chart, suppose that the voltage distribution on a measuring line is as shown in Fig. 72a and the impedance of the termination is to be computed. The wavelength on the line is found from the distance between the two adjacent short-circuit minimums to be 50 cm and the distance from the short-circuit minimum at B to the minimum on the voltage distribution for the unknown impedance at A is $\Delta w = 8.13$ cm. The VSWR for the unknown termination is 2.6:1. Enter the chart at point A (Fig. 72b) which corresponds to a voltage minimum on a 2.6:1 VSWR distribution. Move counterclockwise around the chart (toward load) on the 2.6:1 circle through an angle $2\beta\,\Delta w = 4\pi\,\Delta w/\lambda$. This angle can be found by locating the point $\Delta w/\lambda$ on the scale along the circumference of the chart. Since $\Delta w/\lambda = 0.1626$ the new point is at B. The value of

$$z_1 = r_1 + jx_1$$

as read from the chart at point B is the normalized impedance of the termination. In this case it is $r_1 = 1 - j$. The actual impedance of the termination is

$$Z = R_c(r_1 + jx_1)$$

It should be noted that, if the short-circuit minimum at point C is used as a reference, then the 2.6:1 circle is traversed in the clockwise direction (toward the generator) for a distance corresponding to $\Delta w/\lambda = 0.3374$ which again leads to point B.

References

1. Arguimbau, L. B.: "Vacuum Tube Circuits," p. 348, Wiley, 1948.
2. Terman, F. E., and J. M. Pettit: "Electronic Measurements," 2d ed., Chap. 1, McGraw-Hill, 1952.
3. Honnell, M. A.: An Electronic Potentiometer, *Proc. IRE*, **30** (10), October, 1942.
4. Kip, A. F., A. G. Bousquet, R. D. Evans, and W. N. Tuttle: Design and Operation of an Improved Counting-rate Meter, *Rev. Sci. Instruments*, **17**, 323, 1946.
5. Bagley, A. S.: The High-speed Frequency Counter, *Hewlett-Packard J.*, **2**, January, 1951.
6. Clapp, J. K.: Frequency Measurements by Sliding Harmonics, *Proc. IRE*, **36** (10), October, 1948.
7. Montgomery, C. G.: "Technique of Microwave Measurements," Vol. 11, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1947.
8. Arguimbau, L. B.: Wave Analysis, *Gen. Radio Experimenter*, **7**, 12, June, 1933.
9. Terman, F. E., R. R. Buss, W. R. Hewlett, and F. C. Cahill: Some Applications of Negative Feedback with Particular Reference to Laboratory Equipment, *Proc. IRE*, **27**, 649, October, 1939.
10. Kretzmer, E. R.: Measuring Phase at Audio and Ultrasonic Frequencies, *Electronics*, **22**, 114, October, 1949.
11. Wind, Moe: "Handbook of Electronic Measurements," Vol. 1, Chap. 4, Polytechnic Institute of Brooklyn, 1956.
12. Terman, F. E., and J. M. Pettit: "Electronic Measurements," 2d ed., Chap. 3, McGraw-Hill, 1952.
13. Thurston, W. R.: A Direct-reading Impedance-measuring Instrument for the U-H-F Range, *Gen. Radio Experimenter*, **24**, 1, May, 1956.
14. Byrne, J. F.: A Null-method for the Determination of Impedance in the 100–400 Mc Range, *Proc. Natl. Elec. Conf.*, **3**, 603, 1947.
15. Fong, A.: Direct Measurement of Impedance in the 50–500 Mc Range, *Hewlett-Packard J.*, **1**, 1, April, 1950.
16. King, D. D.: Impedance Measurements on Transmission Lines, *Proc. IRE*, May, 1947.
17. King, D. D.: "Measurements at Centimeter Wavelength," Chap. 6, Van Nostrand, 1952.
18. King, R. W. P.: "Transmission Line Theory," McGraw-Hill, 1955.
19. Smith, P. H.: Transmission Line Calculator, *Electronics*, **17**, 29, January, 1939.

CHAPTER 15

AUDIO-FREQUENCY AMPLIFIERS

By Richard C. Hitchcock[1]

GENERAL REQUIREMENTS

1. General. The term audio includes some or all of the frequencies from 20 to 20,000 cps. This range contains practically all the fundamentals and most of the harmonics of speech and music. The term audio applies equally well to acoustic waves and to electrical impulses. An audio (a-f) amplifier utilizes low-energy electrical frequencies at its input and delivers the same frequencies at higher energy at its output. An input transducer is usually required to provide appropriate electrical impulses. Typical transducers are microphones, phonograph pickups, tape-head pickups, and radio detectors. At the output, another transducer is needed such as a loud-speaker or headphone to change the amplified electrical impulses into acoustic energy.

2. Voltage and Power Gain. A multistage tube amplifier consists of voltage amplifiers at the input, utilizing high-gain tubes such as triodes or pentodes, and of power amplifiers at the output stage utilizing low-impedance triodes or beam-power tubes. The individual stages are usually *RC*-coupled to each other, although transformers are often employed for this purpose. The power stage is coupled to the output transducer by means of an impedance-matching transformer. Some special circuits are available which make this output transformer unnecessary (Sec. 49).

Single-tube outputs are usually operated as class A, push-pull stages as class AB_1 or class B.[2] Class B amplifiers have high efficiency, delivering relatively high audio output per watt of d-c power input. The d-c power for all amplifiers is usually supplied by a combination of transformer, rectifier, and filter.

3. Design Parameters. The following items are important in choosing or designing an a-f amplifier, the numbers referring to sections that follow:

4. Output power needed
5. Input available
6. Range of frequencies desired, or needed
7. Distortion allowable
8. Stability
9. Power supply
10. Cost
11. Space
12. Service

[1] Syntron Company, Homer City, Pa.

[2] A class A amplifier operates in such a manner that the output wave form for a single tube and any kind of output impedance is substantially the same as the input wave form. Plate current flows for the full 360 electrical degrees of the input emf. When the grid of a class A amplifier is not driven positive with respect to the cathode, it is designated as class A_1. When the grid is driven positive, it is designated as class A_2. Class A_2 is seldom used because of high plate losses. Hence class A usually means that the grid is not driven positive.

A class AB amplifier is operated with sufficient grid bias for a given d-c plate potential that plate current flows for more than 180 deg but less than 360 deg of the input emf. This mode of operation requires two tubes in push-pull in order that the output wave form be nearly like the input wave form. When the grids of the tubes are not driven positive with respect to the cathodes, the designation is class AB_1. When the grids are driven positive, the designation is class AB_2.

A class B amplifier is operated with sufficient grid bias to reduce the plate current almost to zero when no input voltage is applied. Plate current flows for only 180 deg of the input cycle. It requires two tubes in push-pull to produce an output wave form that is nearly like the input wave form.

4. Output Power Required. A few milliwatts of electrical output power is sufficient for hearing aids or personal portable radios; for table radios, television sets, or tape recorders from ½ to 2 watts is necessary. For home high-fidelity sets, 10 to 20 watts is common. Amplifiers with rated outputs over 20 watts are used in industrial plants, theaters, hotels, hospitals, and the like.

To measure the electrical output power from an audio amplifier is relatively straightforward. But this is seldom a complete answer to a given application because a loud-speaker is needed to change the electrical watts to acoustic watts. The needed factor, the efficiency, is the ratio of the acoustic watts to the electrical watts. Unfortunately this cannot be specified for the complete audible spectrum, nor is the human ear equally sensitive to all frequencies. The variation in ear sensitivity is detailed by the Fletcher-Munson curves.[1] Furthermore a given loud-speaker unit, when used in various enclosures, may range in efficiency over a range of 10 to 1.

As approximate working figures, *the acoustic power* required for a small apartment is about 2 watts, for a large living room 4 watts, for a small auditorium 8 watts. For

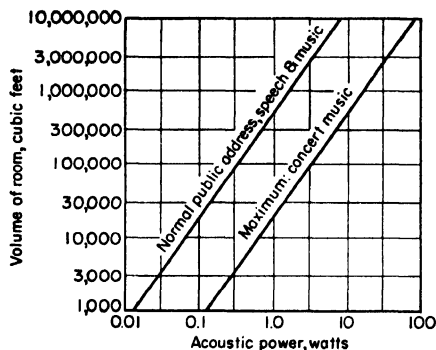



FIG. 1. Indoor acoustic power requirements. (*From Western Electric.*)

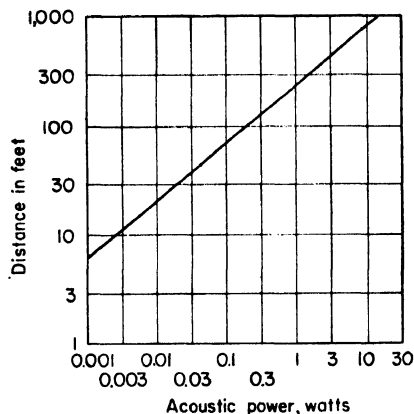



FIG. 2. Outdoor power requirements for speech. (*From Western Electric.*)

home use it is always possible to "turn it down," and so extra electrical power or more efficient speakers may be employed. However, for a multispeaker system, such as for industry, it is essential that as much planning go into the system as possible to ensure its adequacy.

As a rough guide, a loud-speaker mounted on a flat baffle, or in an open-backed box, has an efficiency of 5 per cent; in a closed padded box 10 per cent; in a bass-reflex cabinet 20 per cent; and in an exponential horn, up to 40 per cent. Small exponential horns are used for speech reinforcement and public-address work, although their frequency response is limited. Their efficiencies in the speech range may be as high as 25 per cent. Some commercial speaker enclosures exchange efficiency for smooth and wide-range response. One such device[2] has an acoustic efficiency of 1 per cent but it has only 1.7 cu ft inside volume.

The output power of electrical energy needed, therefore, depends on both the speaker efficiency and the space to be supplied with acoustic energy. A 20-watt amplifier with a 5 per cent efficient speaker will supply 1 watt of acoustic energy. This is somewhat low for high fidelity. A 20-watt amplifier with 25 per cent efficient

[1] Fletcher, H., and W. A. Munson, Loudness, Its Definition, Measurement and Calculation, *J. Acoust. Soc. Am.*, **5**, 82, October, 1933. Reproduced p. 826 in F. Langford-Smith, "Radio Designer's Handbook," 4th ed., RCA, 1953. Also Fletcher, H., "Speech and Hearing in Communication," 2d ed., p. 84, Van Nostrand, 1953. See Chap. 1, Fig. 10, or Chap. 11, Fig. 8, of this handbook.

[2] Acoustic Research AR-1 speaker system.

small horns will put out 5 watts of acoustic energy (with restricted bass response), enough for a space 100,000 cu ft, according to Fig. 1.

The chart in Fig. 1 shows the acoustic watts required for indoor enclosures from 1,000 to 10,000,000 cu ft, both for public-address speech, and (multiplied by 10) for the maximum needed for concert music. For outdoor work, Fig. 2 shows the acoustic watts for 10 to 10,000 ft distance, over an angle of 30 deg. Multiple speakers are needed if the angle is greater than this.

As a rule, the loud-speaker load is closely matched to the amplifier output impedance. This is essential for maximum fidelity and power output. However, for public-address and similar work, it is often close enough to use transducer loads from 75 to 125 per cent of the amplifier rated output impedance.[1]

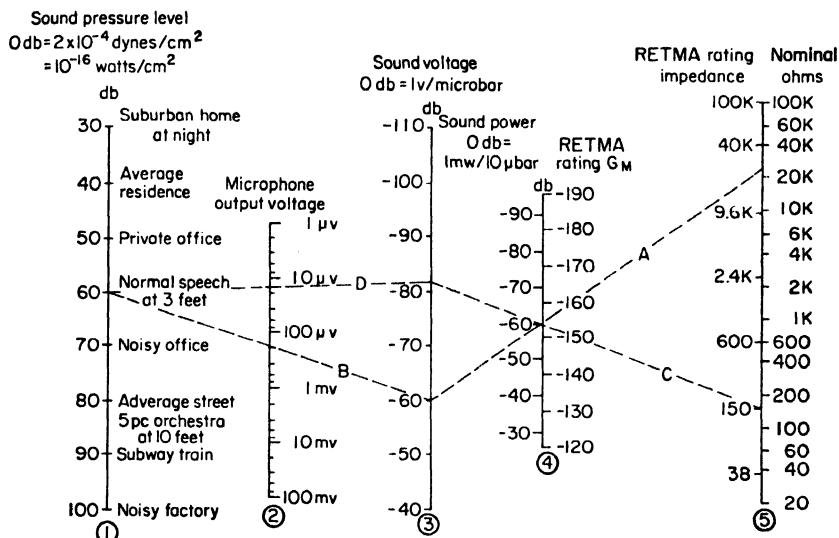



FIG. 3. Microphone ratings. *A-B*, typical high-impedance; *C-D*, low-impedance.

5. Inputs Required. In Table 1 typical ranges of output rms volts and load impedance are listed. Further mention of the transducers noted there will be made below.

Microphones. Microphone ratings are shown in Fig. 3.[2] Scale 3 is the sound-voltage rating ordinarily specified for high-impedance microphones. An example will illustrate the use of these data.

A dynamic moving-coil microphone with its built-in stepup transformer is rated at -60 db sound voltage, scale 3. This is enough information to find the desired microphone voltage on scale 2, but to show the relationship to sound power, draw dashed line *A* to the nominal impedance, scale 5, at 25,000 ohms. This microphone has a -60 db sound-power rating on scale 4, where 0 db is 1 mw per 10 microbars (10 dynes/cm^2). Line up -60 db sound voltage, scale 3, with the sound-pressure level 60 db on scale 1, showing the microphone, output voltage of 200 μv on scale 2, dashed line *B*.

This microphone also has an optional low-impedance connection of 150 ohms, scale 5, with -60 db sound power, scale 4. Dotted line *C* intersects scale 3 at -82 db sound voltage; this intersection is aligned, using dotted line *D*, with the same initial sound-pressure level on scale 1, to show a microphone voltage of 16 μv on scale 2. For scale 3 values below 100 μv a stepup transformer is used, either built into the case or connected at the end of a

[1] White, Saul J., How Far Can I Mismatch? *Audio Eng.*, **35**, 15, January, 1951.

[2] Conover, W. B., Microphone Nomogram, *Electronics*, October, 1955, p. 172. Rosenman, Leo, Microphone Sensitivity Conversion Chart, *Trans. IRE*, PGA AU-1, 33, March–April, 1953.

low-impedance shielded microphone line. Low-impedance microphone lines are desirable since pickup of extraneous signals is minimized, and the high frequencies are attenuated less owing to cable capacitance.

Table 1. Input Transducers

	Output, volts rms	Load impedance, ohms
Ceramic and crystal phono pickups	0.5–2.0	500,000–3,000,000
Reluctance (magnetic) phono pickups	0.005–0.05	10,000–50,000
Microphones, high-impedance	0.0002–0.02	25,000–100,000
Tape-head pickups	0.005–0.01	200–3,000
Vibration pickups	Various	Various
Radio and TV detectors	0.5–5.0	50,000–500,000

Crystal microphones invariably are high-impedance, 25,000 to 50,000 ohms being common. Some microphones have an adjustable response to accentuate the higher-frequency consonants in speech sounds, as much as 10 db.[1] By exposing both surfaces of the diaphragm, the response of another microphone type is better to close-up sounds than to far-away sounds.[2]

Phonograph Pickups, Tape-head Pickups. These will be treated in more detail in Sec. 17.

Vibration and Noise Pickups. Sound-level meters and vibration pickups are a specialized field, and several commercial devices are available.[3] The signal for which an analysis is desired may be, for example, a noise, a complex musical sound, or a mechanical vibration.[4] An analyzer is usually a combination of a filtering system, an amplifier, an attenuator, and an indicating meter. Often these units are combined in a single unit that cannot be tested or rated separately.

Pressure transducers of barium titanate are[5] available with ranges of 0.5 to 30,000 cps flat to ±2 db and with sensitivities of −101 db referred to 1 volt per microbar. These will withstand static pressures of 5,000 psi and operate at −20 to 90°C.

One vibration meter[6] is battery-powered, transistor-operated, etched-circuit-wired, and has a sensitivity of 0.005 in. peak to peak, giving full-scale deflection from 50 to 1,000 cps and usable down to 10 cps.

6. Frequency Range. For pleasing reproduction of music it is an experimental fact that the upper and the lower range of frequencies have a definite relation.[7] If all frequencies are equally amplified and reproduced, the top and bottom frequencies multiplied together should give a figure between 400,000 and 500,000. Taking the lowest figure, 400,000, three systems may be mentioned: fair, 80 to 5,000 cps, 6 octaves; good, 40 to 10,000 cps, 8 octaves; and excellent, 20 to 20,000 cps, 10 octaves. The "fair" system is characteristic of ordinary home radio receivers and phonographs.[8]

The over-all frequency response of speech-reinforcing systems covers only a restricted range[9] but should be at least from 400 to 4,000 cps, and preferably one octave

[1] Electrovoice Cardax model 950, crystal microphone.

[2] Beaverman, W. A., and A. M. Wiggins, A Second Order Gradient Noise Cancelling Microphone, Using a Single Diaphragm, *J. Acoust. Soc. Am.*, September, 1950.

[3] General Radio Co., 760B sound analyzer, 761A vibration meter, 1550A octave band noise analyzer, 1552B sound-level calibrator, 1555A sound-survey meter, 1551A sound-level meter, 762B vibration analyzer.

[4] Acoustic Society of America, "Specifying the Characteristics of Analyzers Used for the Analysis of Sounds and Vibrations," Z24.15-1955, Feb. 4, 1955; American Standards Association, Inc.

[5] Model BG-50, Atlantic Research Corp.

[6] Type 1-128 meter with 4-120A pickup (vibration meter); Consolidated Electrodynamics Corp.

[7] "Frequency Range and Power Considerations in Music Reproduction," Technical Monograph No. 3, 7th printing, 1950, Jensen Mfg. Co.,

[8] Morgan, H. K., Natural Sound Reproduction, *Trans. IRE (Audio)*, **AU-2**, 106, July–August, 1954.

[9] Snow, W. B., Frequency Characteristic of a Sound-reinforcing System, *J. Audio Eng. Soc.*, **3**, 74, April, 1955.

further at each end, or from 200 to 7,000 cps for auditoriums of 237,000 to 5,000,000 cu ft.[1]

Interesting experiments are being performed to show how much speech can be compressed either as a function of frequency range or in the total time required to present it, but these tests are not considered here.[2]

7. Allowable Distortion. All amplifiers have distortion, that is, they are not true linear devices. The important matter is to determine what kind and how much distortion can be permitted for a given application. The requirements placed on an amplifier for a restricted-range small radio receiver are altogether different from those needed in a home-type high-fidelity system.

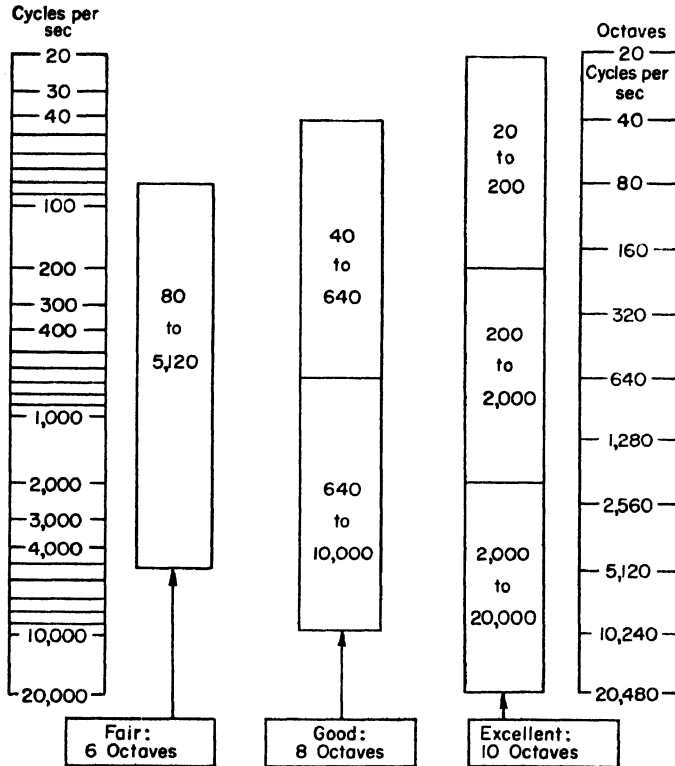



FIG. 4. Audio-frequency ranges required for various types of service.

In general, amplifier distortion is of two kinds. One type is harmonic distortion produced by nonlinearity and composed of higher harmonics of the input voltage. The other type, known as intermodulation distortion (IMD), is evidenced by the modulation of one frequency by another and by the production of sum and difference frequencies not present in the input to the amplifier. IMD is considered as the more important in high-quality systems. Table 2 gives an indication of the values of frequency range, power output, and distortion likely to be found in several audio systems.

[1] Beranek, L. L., W. H. Radford, J. A. Kessler, and J. B. Wiesner, Speech Reinforcement Evaluation, *Proc. IRE*, **39**, 1401, November, 1951.

[2] Fairbanks, G., W. L. Everitt, and R. P. Jaeger, Method for Time or Frequency Compression of Speech, *Trans. IRE (Audio)*, **AU-2**, 7, January–February, 1954.

Kock, W. E., Speech Band Width Compression, *Bell Lab. Record*, **34**, 81, March, 1956.

Table 2. Typical Audio Amplifiers

	Range, cps	Output power, watts	Total harmonic distortion, %
Telephone:			
Commercial	200–3,000	0.006	
Permanent and temporary AM broadcast, schedule A	50–5,000	0.006	
FM broadcast and TV audio	50–8,000	0.006	
FM broadcast, schedule AA (less frequently used)	50–15,000	0.006	
Portable radios, tubes, and transistors	200–4,000	0.03–0.15	10
Table radio, table TV, portable phono	150–4,500	0.5–2.0	10
Console radio, console TV	70–7,000	2–10	10
Apartment, medium-fidelity phono	50–15,000	2–5	5
High-fidelity phono, large room	20–20,000	10–20	2
High-fidelity phono, auditorium	20–20,000	20–50	2
Public address, speech	300–5,000	10–100	5–10
For reference:			
Human-ear range, average all ages	20–10,000		
Young and exceptional ears	20–20,000		
Standard 88-note piano, fundamentals (A_0 = 27.5; middle C_4 = 261.3; A_4 = 440; C_8 = 4,186)	27.5–4,186		

It has been demonstrated that 5 per cent IMD is too much for a high-quality system, that 2 per cent can be detected by a keen ear. As little as 1 per cent is attainable by good design.[1]

8. Stability. A new amplifier should be tested by using an oscilloscope across the output, to see if inaudible frequencies are present. It is possible for an amplifier completely quiet to the ear to have, say, a 50,000-cycle oscillation, which may overload the output tubes and thus cause poor quality in the audible range.[2]

9. Power Supply. Usually a conventional transformer-rectifier-filter system supplies the d.c. for plates, screen grids, and bias voltages; see Fig. 25. Tube heaters are usually run on a.c., though for high-gain input stages, d.c. is desirable to decrease hum.

For portable work, battery-type tubes with low current drain are necessary. The transistor is excellent for this application. Transistors have no filament power to furnish, but they have at least one complicating difference from vacuum tubes—their inputs and outputs are interdependent, requiring care in design.

10. Cost. Some idea of relative cost may be obtained by considering the various types of components involved in an audio amplifier. The output transformer, when called on to reproduce 8 or 10 octaves, with minimum distortion and faithful response, is the most expensive component. This is one reason for designing amplifiers which do not require an output transformer (see Sec. 49). Beam-power tubes handle more power than triodes or triode-connected pentodes, and to keep costs down beam pentodes are attractive. Good interstage and input transformers are not cheap, though not so expensive as the output transformer because of their lower level of operation. Many amplifiers use resistors and capacitors, some with high-gain pentode tubes, to eliminate the interstage transformers. Microphones with low impedances require a stepup transformer, as mentioned in Sec. 5.

Practically all power amplifiers use a power transformer. An alternative is to connect the heaters of the tubes in series and to connect the "string" across the power line. The warm-up times of the tubes should be equal. Some plate-voltage sources use doublers (Sec. 48). Both these methods will eliminate the need for a power transformer but they add the hazard of line voltage appearing on the metal chassis. To avoid this trouble an isolation transformer between the device and the power

[1] Hitchcock, R. C., Intermodulation Distortion, *Audio Eng.*, **36**, 21, October, 1952. Jones, E. M., How Much Distortion Can You Hear? *Trans. IRE (Audio)*, **AU-2**, 42, March–April, 1954.

[2] Hitchcock, R. C., Testing with White Sound, *Audio*, **38**, 41, October, 1954.

line is sometimes employed but this puts back into the device the costs saved in eliminating the power transformer.

11. Heat Dissipation. A power amplifier must be able to get rid of some of its heat. It is seldom advisable to enclose a power amplifier completely; air circulation must be provided. Amplifier chassis usually have louvers which indicate where such circulation is expected.

12. Service. From time to time, tubes have to be removed for test or replacement. If an enclosure is used, some arrangement should allow tube removal with a minimum of trouble. Other components, which are moved often, such as volume controls, potentiometers, and switches, occasionally have to be replaced. The ready service of components should be designed into the amplifier. Some makers use the "one-layer" construction, so that each component can be reached once the chassis is removed from the cabinet. Others use a "layer-built" scheme. This requires the removal of one component to get at another beneath it.

Components potted in plastic cannot be separated; the complete assembly must be replaced. Printed circuits require special care when components are to be replaced.

13. Feedback. By principles which are now well known, a certain amount of voltage or current may be fed back into the input of an amplifier from the output in such phase as to subtract from the input voltage. By such negative feedback, the amount of distortion and noise produced by the amplifier may be reduced so that it does not appear in the loud-speaker. If the input voltage normally applied to the amplifier must be increased by a factor of 5 (14 db) to secure the same output power obtained without the negative feedback, the intermodulation distortion will decrease by a ratio of 5:1, approximately.

Thus, a 6L6 push-pull amplifier with no feedback will have approximately 5 per cent IMD at 5 watts output and when 14 db feedback (5:1 in voltage) is utilized, the IMD will be about 1 per cent.[1]

Negative feedback will decrease distortion, improve the transient stability, reduce noise, and slightly extend the frequency range. It will not increase the power output. It is most useful where the voltages applied to the amplifier are of the order of 1 volt or more.

In practice, some voltage or current is secured from the output of the amplifier and is applied to a point near the input, say to a single-ended driver stage. The feedback may be secured in parallel with or in series with the load by appropriate measures, sometimes by using a separate feedback winding on the output transformer, sometimes by a voltage-dropping resistor in series with the cathode-bias resistor of the stage to which the voltage is applied, and sometimes by taking the voltage drop across a low resistance in series with the loud-speaker. In the latter case it is known as current feedback and is particularly useful where the loud-speaker has an inadequate enclosure. Current feedback is seldom applied alone but is usually accompanied by some voltage feedback.

The feedback voltage may also be obtained from the primary of the output transformer by use of a resistor to lower the voltage and a capacitor in series to block the d-c voltage. Feedback obtained in this way may be usefully applied when the output transformer is not top quality because the feedback is not affected by leakage reactance.

14. Feedback Design. The principles involved may be readily ascertained by a typical calculation.

In Fig. 5 assume that a 9-ohm load is connected to the 9-ohm tap on the output transformer, that E_0 for 20.6 watts output is 13.6 volts. The input required e_i is 0.19 volt. Feedback is obtained by the voltage divider R-R_f, the portion appearing across R_f being applied across the cathode-bias resistor of the input tube.

Let E_f be the feedback voltage

$$E_f = E_0 R_f/(R + R_f) = 13.6 \times 470/4{,}070 = 1.57 \text{ volts}$$

[1] Hitchcock, R. C., Intermodulation Distortion, *Audio Eng.*, **36**, 21, October, 1952.

The total input voltage now required is

$$E_i = e_i + E_f = 0.19 + 1.57 = 1.76 \text{ volts}$$

Thus, with no feedback, 0.19 volt drives the amplifier to 20.6 watts output; with feedback 1.76 volts is required.

The feedback ratio $E_f/e_i = 1.57/0.19 = 8.26 = 18.3$ db.

Feedback ratios of 5:1 to 10:1 (14 to 20 db) are customary.

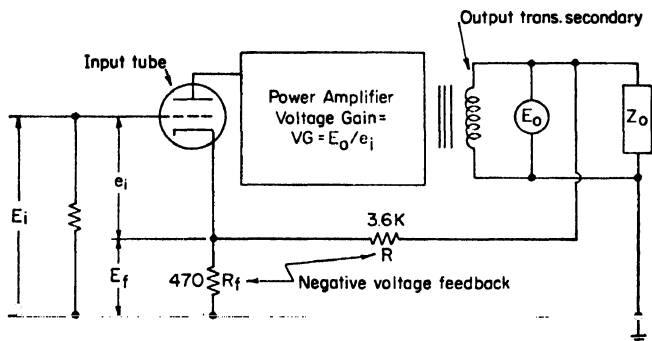



FIG. 5. Feedback from output transformer secondary winding.

PREAMPLIFIERS

Because the output voltage of most transducers is very low, much less than 1 volt, considerable voltage amplification is necessary before the final power amplifier can be driven. In general this preliminary voltage amplification takes place in a separate unit called a preamplifier characterized by high gain and low noise and with whatever tone compensation is to be employed built into it.

The design of these amplifiers is straightforward. Descriptions of several types to follow illustrate the general arrangement.

15. Multiple-input Preamplifier. A preamplifier having five inputs, usable one at a time, designed for home-type high fidelity, is Fig. 6. There are three high-impedance inputs for tuner, television, and tape and two low-impedance inputs for magnetic pickups or microphone. At 1,000 cps for 1.5 volts output the high-level inputs require 0.075 volt, and the low-impedance inputs 0.0015 volt. Each high-level input has a potentiometer to preset the volume of incoming signals. Low-level circuits also have a volume control but effective after the signals have been amplified by the first dual triode. There is sufficient gain and equalization to play directly from a magnetic tape head, using the low-level magnetic pickup jack.

The two input tubes are operated at reduced voltage, and with d.c. on the heaters. The noise level is 95 db below 1.5 volts output. Two outputs, cathode-follower type, are provided.

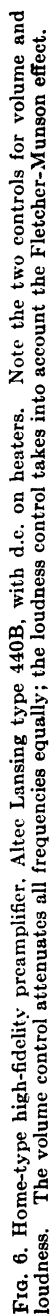
Selector switches determine low-frequency turnover (bass boost compensation begins) and high-frequency rolloff (treble attenuation begins), with a total of 25 different positions. In addition the bass control can supply +13 db or −9 db at 100 cps, and the treble control at 5,000 cps can furnish +10 db or −9 db.

Plate current is supplied by two metal rectifiers and an *RC* filter, independent of the rectified heater supply also using metal rectifiers.

The tubes may be removed from the back without removing the chassis enclosure. Three power receptacles are energized by the on-off switch, for tuner main amplifier, phonograph motor, for example.

16. Single-input Preamplifier.[1] A preamplifier designed solely for magnetic pickups (Fig. 7) using two 6SJ7 pentode amplifiers and a 6SL7 double cathode follower provides 2 volts output. A voltage-regulator tube supplies a constant voltage needed for 40-db feedback. Any amplifier which has high gain at low frequencies is susceptible to "motor-

[1] Finlay, Alexander, A Simple—Yet Good Quality—Preamp, *Radio and Television News*, **52**, 42, November, 1954.

FIG. 6. Home-type high-fidelity preamplifier, Altec Lansing type 440B, with d.c. on heaters. Note the two controls for volume and loudness. The volume control attenuates all frequencies equally; the loudness control takes into account the Fletcher-Munson effect.

boating," and the use of the voltage-regulator tube isolates the cathode-follower plate circuits from the others. An external source of 250 volts d.c. and 6.3 volts a.c. is required. The 6SJ7 cathode resistors are lower than normal to obviate the need for bypass capacitors and allow maximum transconductance.

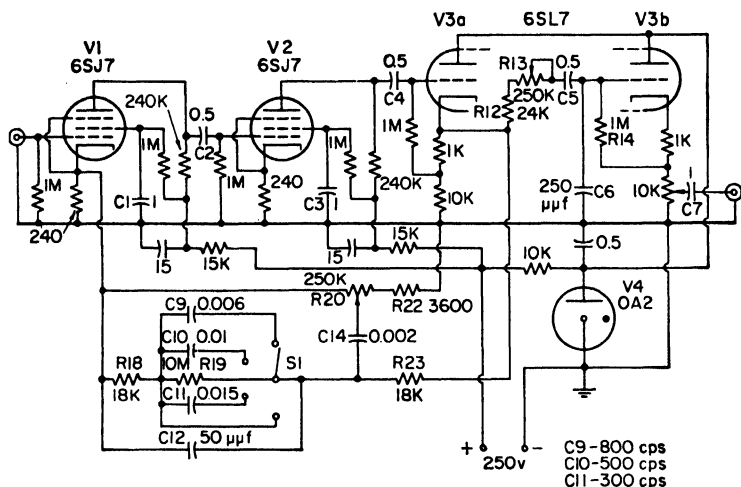



FIG. 7. Preamplifier for use with magnetic pickup.

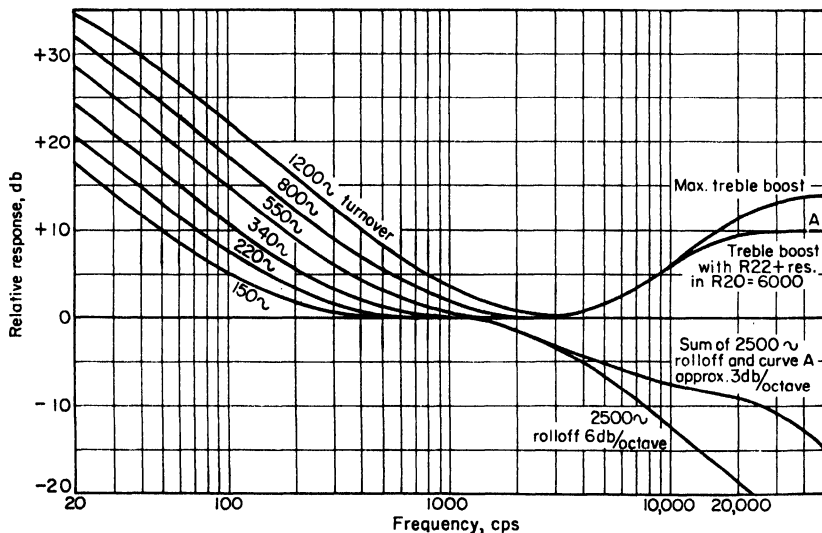



FIG. 8. Preamplifier equalization curves.

The feedback loop comes from the first cathode follower by way of R_{23}, switch S_1 to the cathode of the input 6SJ7. This circuit shunts C_9, C_{10}, and C_{11} across R_{19}; the fourth switch point shunts C_{12} across R_{19} and R_{18} for the highest turnover frequency shown in Fig. 8. No bass boost is provided, but much the same effect can be obtained by increasing the turnover frequency. At 150-cps turnover, for example, the boost at 100 cps is 5 db, and with a 1,200-cps turnover, the boost at 100 cps is 22 db. Other turnover frequencies may

be obtained by using a series capacitor to replace C_9 in the feedback loop, with a value of $\frac{1}{4}\pi f_T R_{18}$, where f_T is the new turnover frequency and R_{18} is the 18,000 ohms (R_{18}) to the first 6SJ7 cathode.

A hum balancer across the heater winding, which puts 20 volts positive d.c. on the variable arm, will reduce hum. Substituting type 5693 tubes for the 6SJ7s makes this unnecessary. High-frequency rolloff is controlled in the output cathode follower V_{3b} grid circuit by R_{14}, C_6, and R_{13}. Different values may be obtained for a rolloff frequency f_R by changing R_{13} plus R_{12} to a value given by $\frac{1}{2}\pi f_R C_6$, where C_6 is 250 × 10^{-12} farad.

Treble boost is the R_{20}, R_{22}, C_{14} circuit from the first 6SJ7 cathode to ground. Set at mid-position, the response is flat; at one extreme of R_{20} half the feedback loop is shunted by C_{14}, giving a reduction of 6 db above 6,000 cycles. At the other extreme, there is an increase of amplification above 4,000 cycles.

Higher-value capacitors C_1, C_2, C_3, C_4, and C_5 contribute to the good low-frequency response. Tests showed sine-wave response from 1 to over 300,000 cps. Square-wave tests showed negligible tilt at 4 cps; at 1,200 cps there was no overshoot or ringing. Intermodulation distortion tests showed "no deflection of the meter on a 3 percent full-scale range."

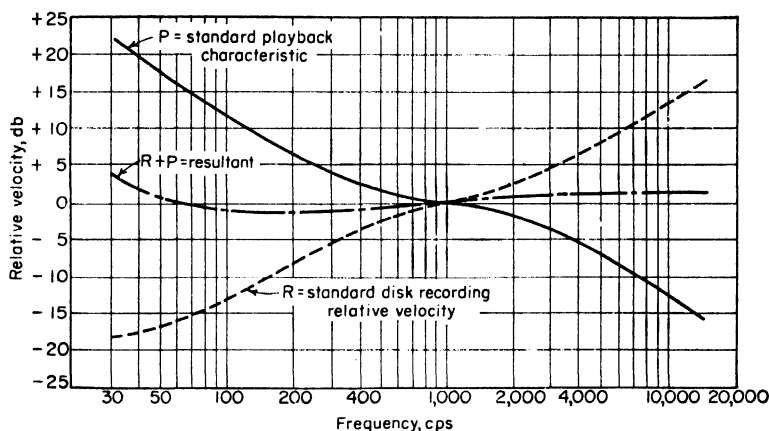



FIG. 9. Characteristics of disk recording and playback.

17. Preamplifier Frequency Compensation. The frequency characteristics used during the recording of a phonograph disk determine what the response of the playback pickup must be to get uniform output for all frequencies. During recording, lateral velocities are limited more and more as the frequencies go lower and there is a limiting maximum lateral excursion of the stylus which will use the record space efficiently for both medium and low frequencies. At the other end of the spectrum, there is a preemphasis during recording, to improve the signal-to-noise ratio.

The "standard" commonly used by recording-disk manufacturers in the United States is the dashed line of Fig. 9. At 100 cps the relative velocity is −13.1 db, and at 10 kc it is +13.7 db.[1]

To counteract this characteristic the standard playback is the solid line of Fig. 9 which shows +12 db at 100 cps and −12 db at 10 kc.* The resultant is shown by the dot-dash line; +4 db at 30 cps, 0 db at 70 cps, 0 db at 1,000 cps, and +2 db at 8,000 cycles.

The object of a preamplifier compensation circuit is to provide this solid-line response, when using pickups on records, to feed an otherwise "flat" amplifier. Two such circuits are shown in Figs. 10 and 11.

A commercial preamplifier used for reluctance pickups is shown in Fig. 12. The output

[1] "New Orthophonic Recording Characteristics, RIAA Standard," RCA Record Division. Moyer, R. C., Evolution of a Recording Curve, *Audio Eng.*, **37**, 20, Fig. 4, July, 1953.

* AES Standard Playback Curve, *Audio Eng.*, **35**, 22, January, 1951.

is about 0.5 volt. The dotted input resistor R_1 can be any value from 6,800 to 58,000 ohms; the higher value gives better high-frequency response, but the pickup circuit will then be more susceptible to stray fields which may cause hum.

An inspection of the tone-control curves—Fig. 8 with maximum bass boost, and Fig. 8 with 340-cps turnover—will indicate that relatively flat response can be obtained with these controls without the need of added compensation. However, magnetic pickups have a low output voltage, in the range of 5 to 40 mv, and some amplification is required to reach the level where the tone control can be used.

A typical tape-head response is Fig. 13. This characteristic shows the same decreasing output with decreasing frequency as shown by the dotted line for disk records in Fig. 9. Note the position of 0 db at 200 cps for the tape head: it can be compared with the disk

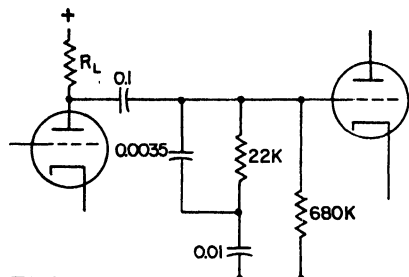



FIG. 10. Triode equalization for standard disk recording.

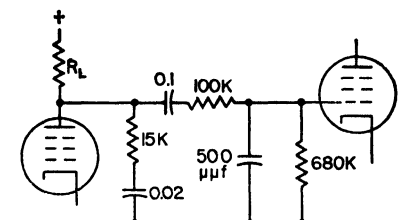



FIG. 11. Pentode equalization for standard disk recording.

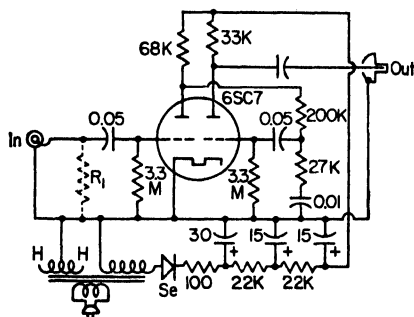



FIG. 12. Compensated preamplifier—General Electric UPX-003.

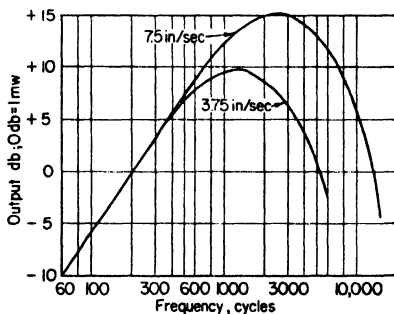



FIG. 13. Tape-head output. Brush Type BK-1090.

record by subtracting 8 db from the tape-head response. Thus the tape shows −18 db at 60 cps, compared with −13 db for the disk; at 1,000 cps the 7½-in.-per-sec tape has +5 db compared with 0 db for disk, and at 3,000 cps the tape is +7 db compared with +5 db for disk. Thus a 5-db bass boost at 60 cps is needed to compensate for the tape recording and still use the same compensation circuits shown for the disk.

18. Tone Controls. The use of a tone control is a controversial subject. It must be recognized that amplifiers, speakers, and input circuits are not perfect, and that the possibility of changing the frequency response may be justified. The simpler the controls, the better the adjustment to suit a particular situation. Inductor types of tone controls are seldom used at present; *RC* controls of the continuously variable type are favored. Two independent controls are customary: base controls affect the gain from 500 to 1,000 cps and below; treble controls, from 1,000 to 2,000 cps and above. The volume control is not a tone control.

A bass- and treble-control stage is shown in Fig. 14.[1] The two controls are poten-

[1] "RCA Receiving Tube Manual RC-17," June, 1955, p. 285.

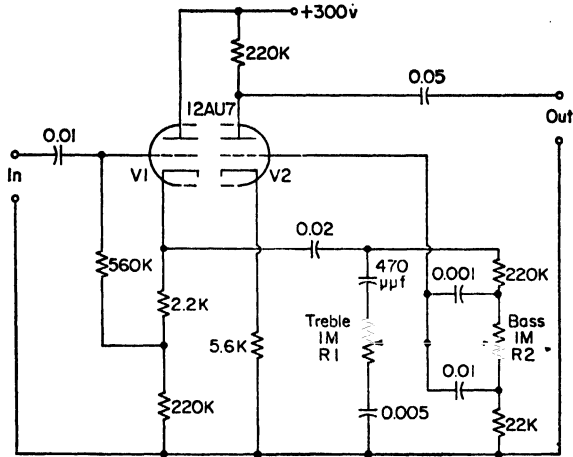



FIG. 14. Circuit for bass and treble control.

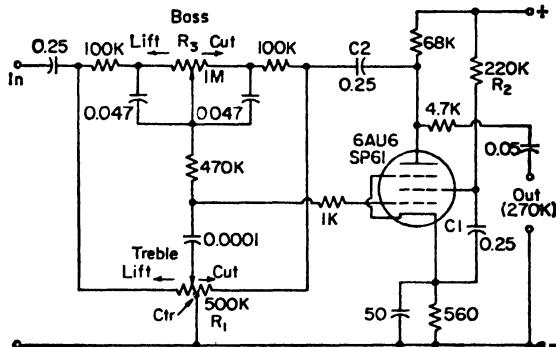



FIG. 15. Feedback type of tone control. With 6SN7 instead of 6AU6, omit R_2 and C_1.

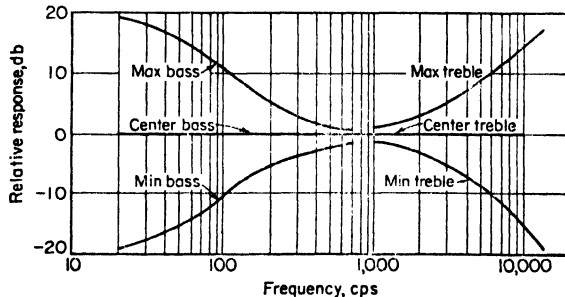



FIG. 16. Tone control available from circuit of Fig. 15.

tiometers R_1 for treble and R_2 for bass, each with logarithmic taper. Boost is obtained as the potentiometer arms are moved "up" and cut as they are moved "down"; flat response is at the middle of the potentiometers. The curves are within 2 db of those shown in Fig. 16. There is no gain from V_1 but V_2 has a voltage amplification of about 10 so that the insertion loss of this circuit is practically zero.

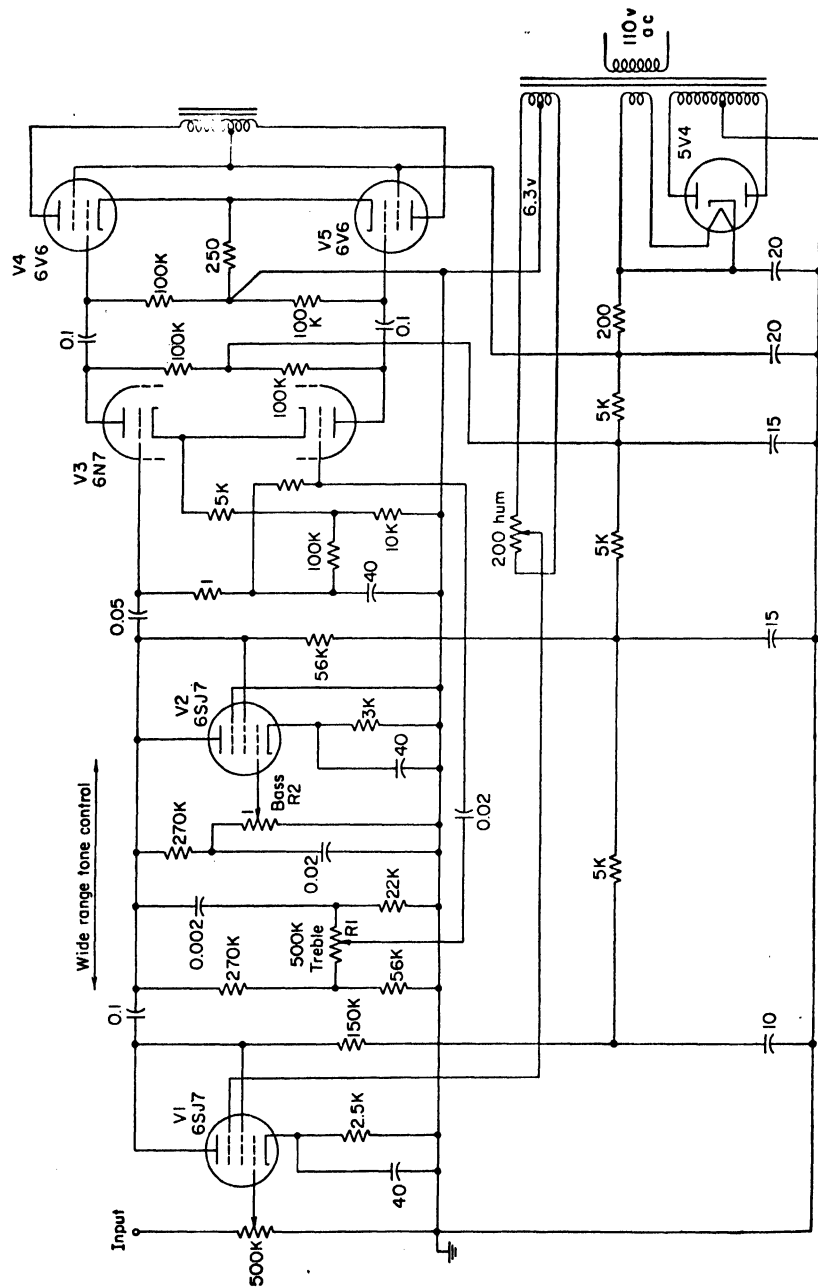



Fig. 17. Wide-range 6-watt amplifier requiring 0.7-volt input.

A negative-feedback tone-control circuit is shown in Fig. 15. Treble control R_1 is a linear potentiometer. Bass control R_3 gives neither boost nor cut at its center. The circuit was originally designed for an Ediswan SP61 tube.[1] A 6AU6 with somewhat lower transconductance may be employed. By using a triode, such as one-half of a 6SN7, and eliminating C_1 and R_2 slight reduction in bass boost will be experienced. The circuit can produce 4 volts rms with less than 1 per cent total harmonic distortion. Negative feedback C_2 provides output essentially independent of the actual tube gain. The mid-frequency gain is approximately unity, so that it can be inserted between a preamplifier and the main amplifier with no loss. The frequency characteristics are shown in Fig. 16.

A wide-range two-channel tone control is shown in Fig. 17.[2] Bass frequencies appear across R_2, are amplified by V_2, and fed to one grid of V_3. When R_1 is at its left-hand position, all frequencies are "flat" and fed by way of the 0.02-μf capacitor to the other 6N7 grid. Treble boost is a maximum of 5 db at 5,000 cps with R_1 arm at its extreme right position. The two halves of V_3 have common cathode coupling, but oppositely phased plates, feeding the push-pull 6V6 tubes.

V_2 is a true bass booster, controlled by R_2, independent of the rest of the amplifier. The bass response ranges from flat, with V_2 grid at ground potential, to +15-db boost at 100 cps with R_2 arm at its top. As indicated in the reference, this circuit is designed for low operating levels, where a large bass boost may be desired. Having no negative feedback, this circuit is less desirable than some. Also, as the treble potentiometer arm leaves its "flat" position there is a corresponding bass cut, which, however, may be filled in by appropriate bass boost.

Note that grid No. 3 of the input tube goes to the arm of a 200-ohm hum adjuster across the 6.3-volt heater supply.

COMPLETE AMPLIFIERS

The majority of these amplifiers are commercially available. They have been selected to show a variety of circuits, or to indicate typical circuit constants.

19. Transistor Hearing Aid. A three-transistor hearing aid called "The Listener" is compactly arranged in a plastic eyeglass frame (Fig. 18) and is powered by a single 1.3-volt mercury cell. The stages are transformer-coupled. R_1 and C_5 in Fig. 19 can be adjusted to change the peak response as much as 10 db in the 1,000- to 3,000-cps range as required. The volume control feeding the second transistor has a range of 40 db. The electrical output to the earphone is 1.0 mw. The semiadjustable ceramic printed resistor R_7 is factory-set for the proper output transistor current.

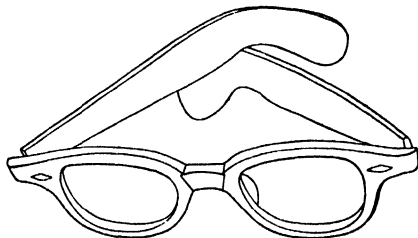



FIG. 18. Spectacle-type hearing aid.

The life per cell is about 10 days with 15 hr use per day with maximum output. The life of the cell is doubled with reduced audio output. A feature is the incorporation of the microphone in the glasses frame (Fig. 18), eliminating the clothing noise of body-worn microphones.

20. Miniature Message Repeater. This is a three-tube recorder-playback, with a maximum of 4 min of recording on a specially processed ¼-in. magnetic tape, started by an external push button. The volume control is in the grid circuit of the second half of the 12AX7 voltage amplifier. The 50B5 output tube supplies 1.2 watts to a self-contained speaker. The microphone is separate, used for the recording and then removed. A 50B5 oscillator provides 38-kc erase power during recording. A half-wave selenium rectifier and *RC* filter supply d-c voltages. Tape speed is 3¾ in. per sec; response ±3 db from 200 to 5,000 cps, which is adequate for speech reproduction in the service intended. See Fig. 21.

21. Table-model High-fidelity Phonograph. This three-tube amplifier (Fig. 22) uses push-pull 50L6 output tubes, with 5 watts maximum output, and a ceramic pickup. The second half of the 12AX7 is a phase inverter. Negative feedback is obtained by a tertiary winding on the output transformer, fed via 3,300 ohms to the first 12AX7 cathode. The feedback is greatest at low frequencies.

[1] Baxandall, P. J., Negative Feedback Tone Control, *Wireless World*, **43**, 402, October, 1952; **43**, 444, November, 1952.

[2] Hill, J. M., Wide Range Tone Control Circuit Suitable for Correction at Low Volume Levels, *Wireless World*, **52**, 422, December, 1946.

Two permanent-magnet speakers provide low-frequency reproduction, and two "kilosphere" electrostatic speakers furnish high frequencies. These speakers are polarized by 200 volts d.c. through two 6.8-megohm resistors. D-c power is supplied by a selenium doubler with a conventional *RC* filter.

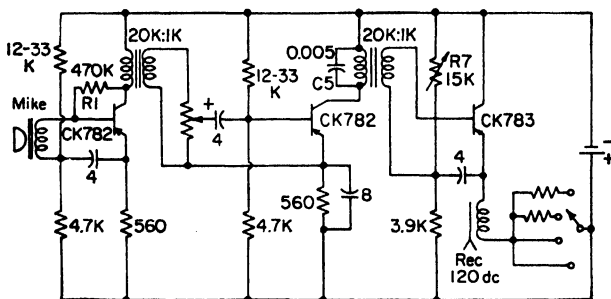



FIG. 19. Transistor hearing aid of Otarion, Inc.

22. Small Public-address Amplifier. A typical small amplifier (Fig. 23) with a single 6L6 output tube has less than 5 per cent distortion at 6 watts output from 400 to 10,000 cps. At 100 cps the 6-watt output distortion is 8 per cent. The input transformer is balanced to ground for a 125- to 600-ohm microphone. Volume controls R_1 and R_7 are for microphone and crystal phono inputs, respectively. Negative feedback to the cathode of the second 6J7 includes network R_{14}, R_8, C_5, R_9, R_{10}. Tone control R_{10} decreases high frequencies as desired.

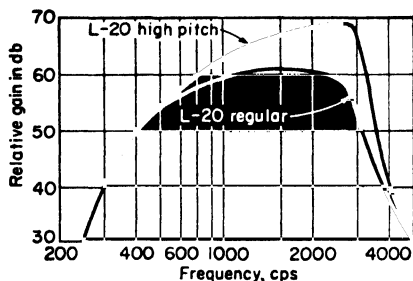



FIG. 20. Over-all response, transistor hearing aid of Otarion, Inc.

23. High-fidelity Amplifier for Home Use. A complete preamplifier power amplifier on a single chassis, with 6L6GB tubes in push-pull output, rated at 22 watts, is shown in Fig. 24. It has bass and treble controls and record-scratch and turntable-rumble filters. The inputs are for two magnetic cartridges, ceramic or crystal pickup, tuner, and tape. Response is 20 to 30,000 cps with harmonic distortion less than 0.8 per cent. Negative feedback is from the 16-ohm secondary winding of the output transformer to the cathode of the 12AX7 push-pull driver via R_{34} and C_{28} in parallel.

24. Scott Dynaural 30-watt Amplifier. This amplifier (Fig. 25) has the dynaural dynamic noise suppressor built in.[1] This is a special type of tone control to eliminate needle scratch on older records. The recorded audio signals must exceed the scratch noise to accomplish this effect. A low-frequency rumble filter is also incorporated.

Three 12AX7 heaters in series are supplied with d.c. from the output tube's bias resistor. Push-pull 1614 beam tubes provide the output with feedback from the 16-ohm output transformer secondary to one of the push-pull 12AX7 driver cathodes. Variable damping is provided, the voltage developed across the 10-ohm resistor depending on the output current drawn. Harmonic distortion is less than 0.5 per cent, IMD less than 0.25 per cent, with output usable from 4- to 24-ohm loads.

25. Williamson Amplifiers.[2] The earlier version (Fig. 26) of these amplifiers was the first to popularize excellent quality for home construction. Some of the features are negligible nonlinear distortion up to maximum rated output, linear response from 10 cps to 70 kc, negligible phase shift in the audible range, good transient response, and low output resistance. It is an all-triode power amplifier requiring 2 volts across 1 megohm for

[1] Scott, H. H., Dynamic Suppression of Phonograph Record Noise, *Electronics*, **19**, 92, December, 1946; Dynamic Noise Suppressor, *Electronics*, **20**, 96, December, 1947.

[2] Williamson, D. T. N., and P. J. Walker, Amplifiers and Superlatives, *J. Audio Eng. Soc.*, **2**, 75, 1954. Williamson, D. T. N., *Wireless World*, **55**, 282, 1949. A booklet, "The Williamson Amplifier," is available for $1 from British Radio Electronics, Ltd., 1833 Jefferson Place, N.W., Washington 6, D.C.

15 watts output. The input 6SN7 is direct-coupled to the phase splitter 6SN7 grid whose plate and cathode in turn feed the output stage. Ten-to-one voltage feedback, 20 db, is taken from the low-impedance loud-speaker terminals and fed to the unby-passed cathode resistor R_4 of the input triode. Feedback resistor R_{25} depends on the output transformer connection used; for 1.7 ohms output R_{25} is 1,500 ohms, for 9 ohms output it is 3,600 ohms, and for 16 ohms output it is 4,800 ohms.

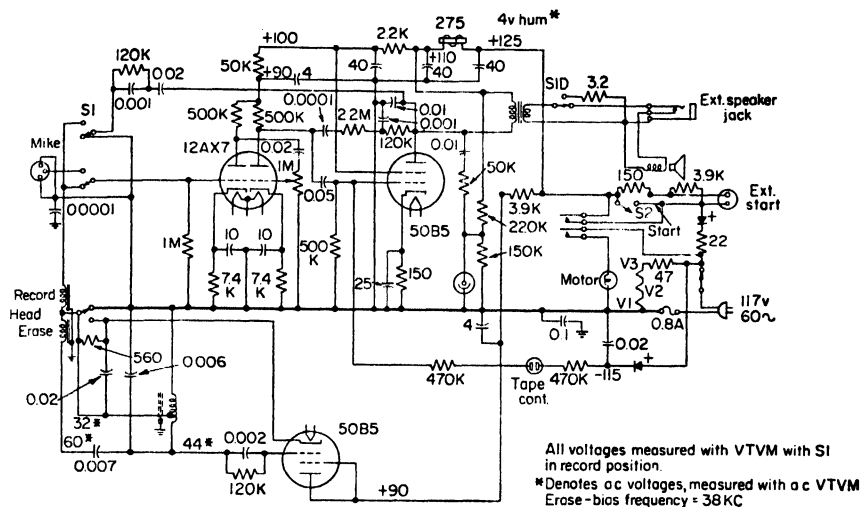



FIG. 21. Circuit of Mohawk "Message Repeater," a dictation machine.

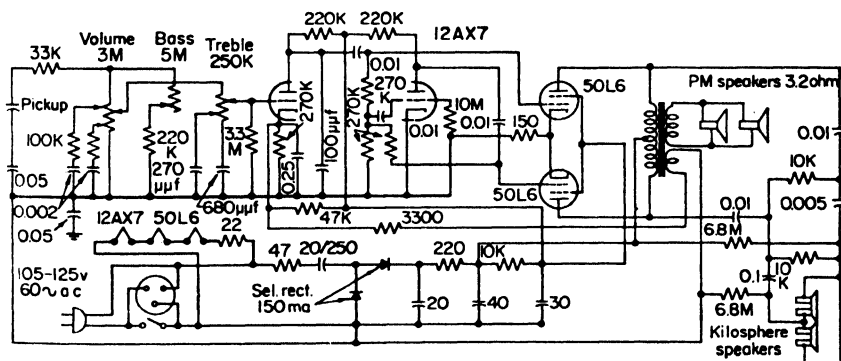



FIG. 22. Table-model home-type high-fidelity amplifier, 5 watts maximum output; two low-frequency speakers, two electrostatic high-frequency speakers (Columbia Records Model 426).

Figure 26*b* gives the frequency range with and without feedback, also the phase shift from 1 to 1,000,000 cps. This amplifier takes negative-feedback voltage around four stages. This is a severe test of design and is one of the outstanding features of the Williamson circuit.

Preliminary adjustments are made by R_{21} to bring the total plate current of the KT66s to 125 ma, and with no signal input, by connecting a 10-volt d-c meter across the plates of these tubes and by adjusting R_{17}, equal plate currents are indicated when the voltmeter reads zero.[1]

[1] Williamson, D. T. N., *Wireless World*, **55**, 282, 1949.

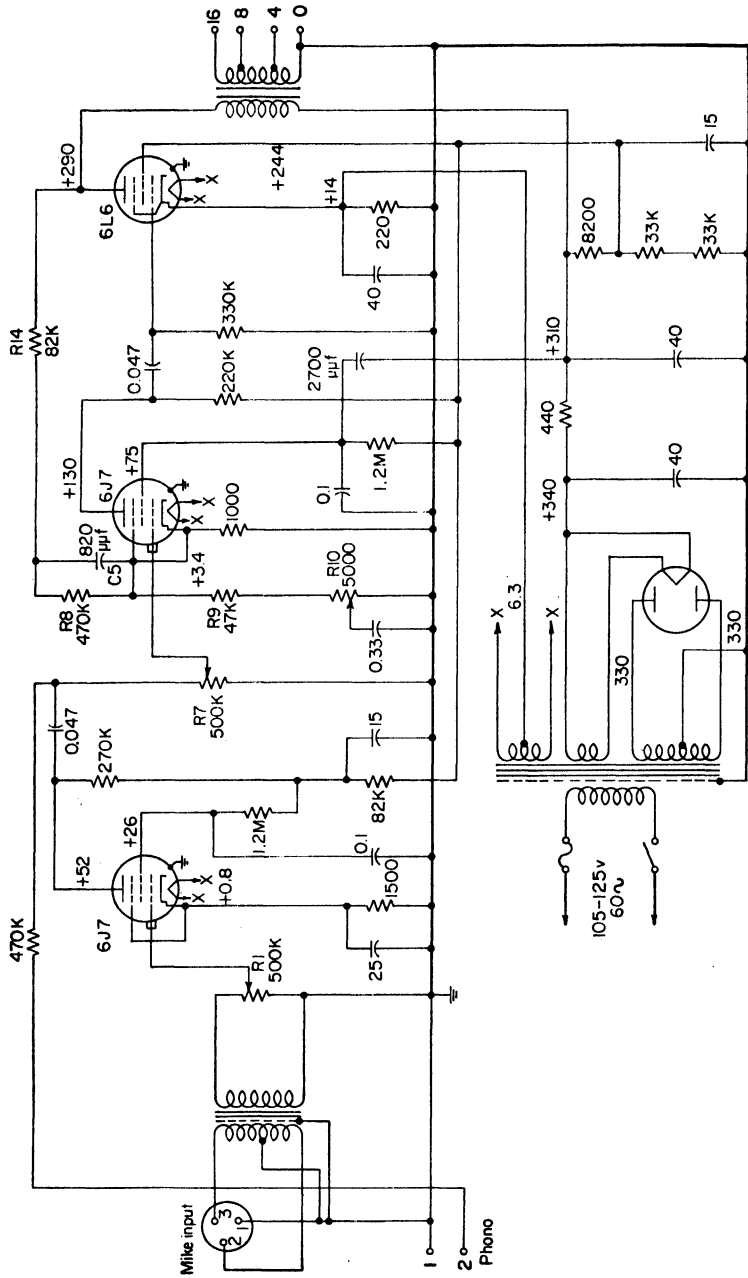



FIG. 23. Small public-address amplifier, single-ended 6L6 output, 6 watts (RCA MI-12238B type SA-6A).

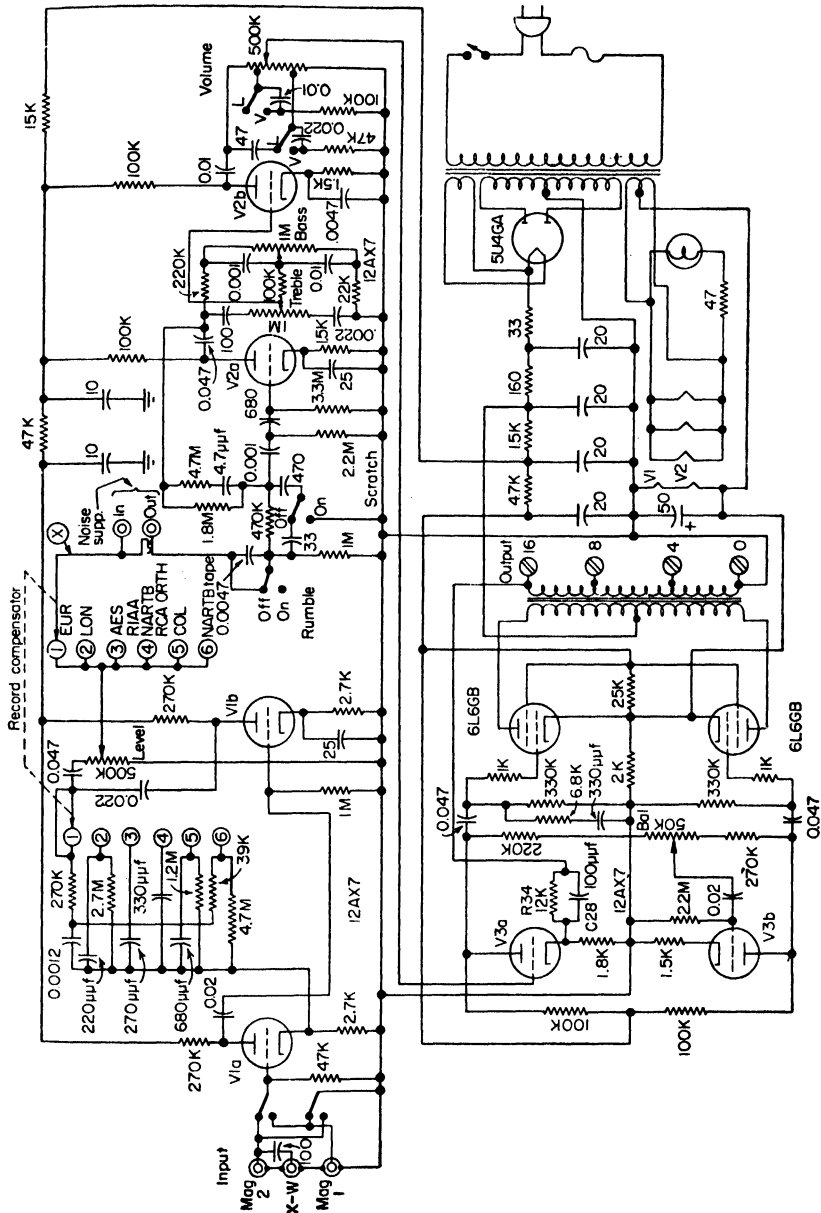



FIG. 24. High-fidelity chassis for home use, 25 watts (H. H. Scott).

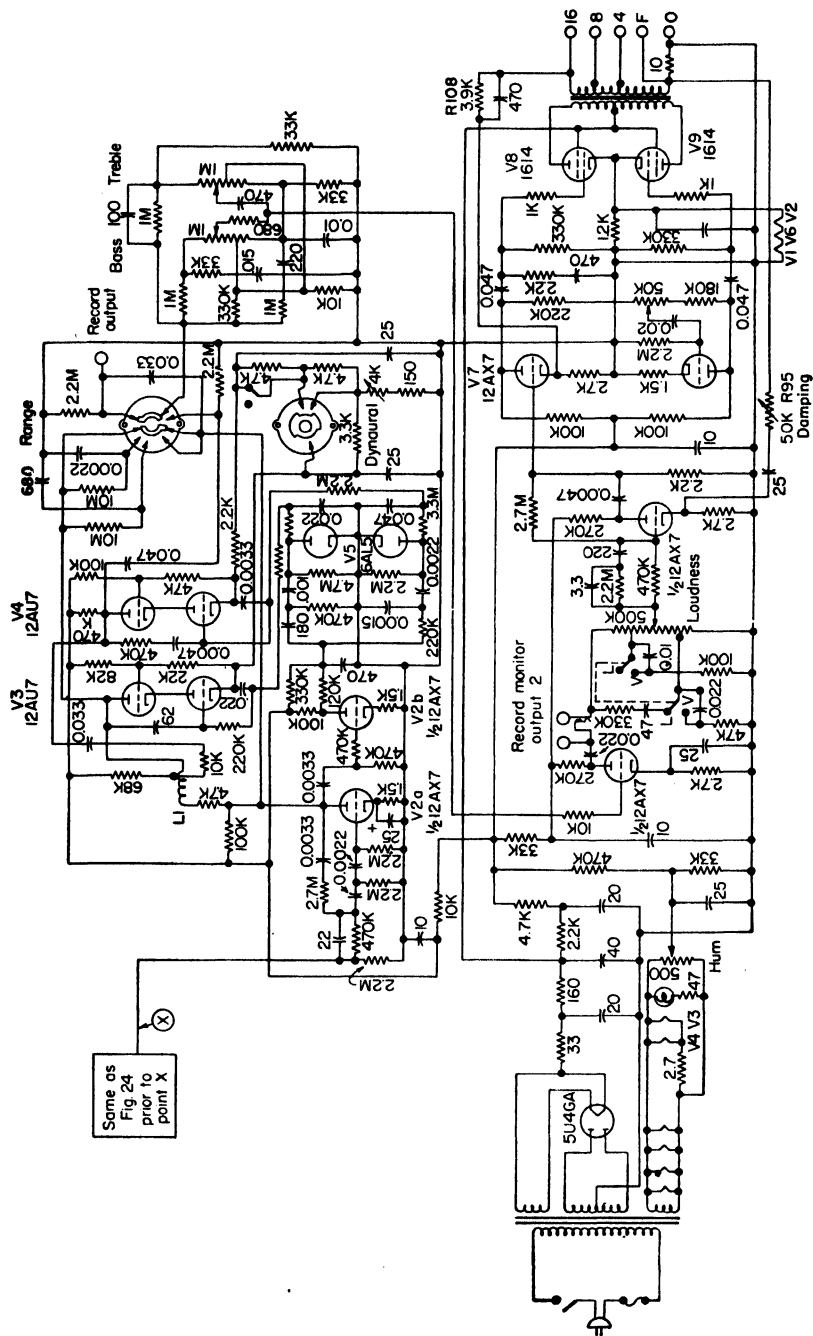



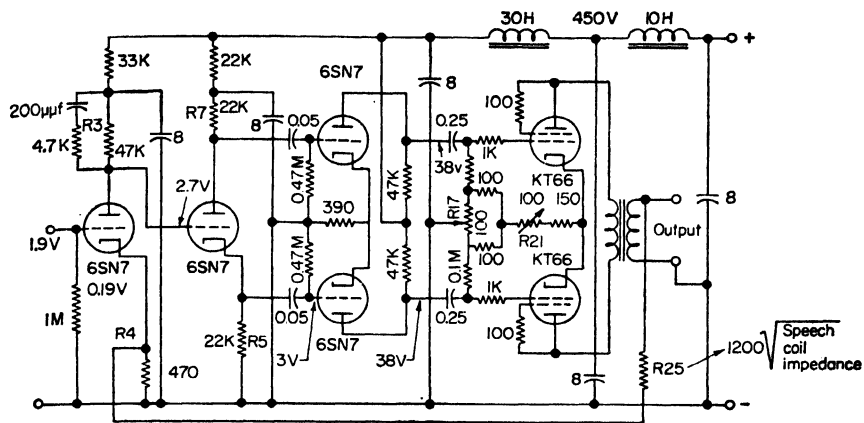
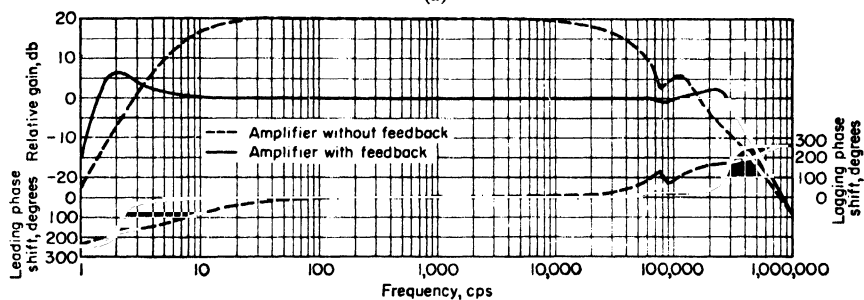
FIG. 25. Scott dynaural amplifier with dynamic noise suppressor and rumble and scratch filters.

(a)

(b)

FIG. 26. Williamson high-fidelity amplifier and loop gain and phase-shift characteristics.

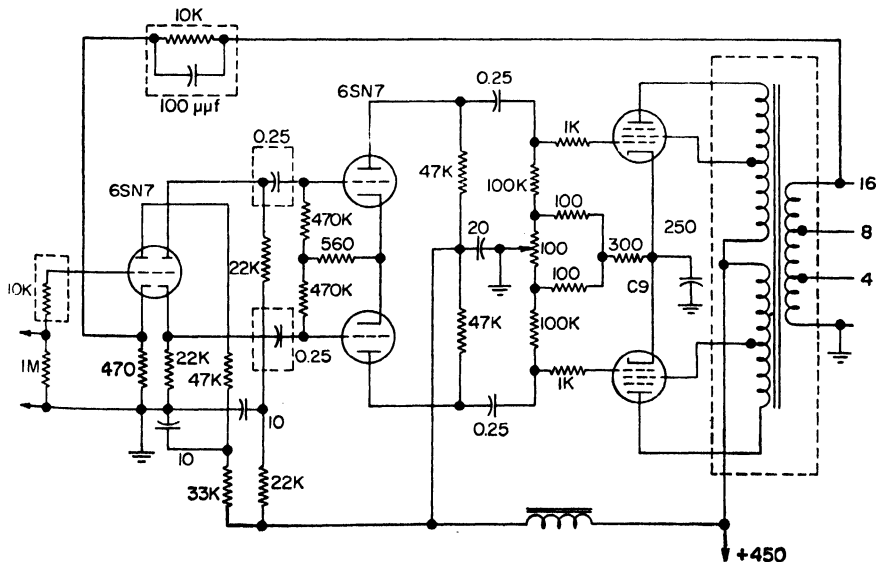



FIG. 27. "Ultralinear" Williamson amplifier.

Ultralinear Williamson.[1] A combination triode-pentode circuit can be used by changing the standard Williamson circuit as shown in Fig. 27. The important change is the output transformer, which may be used with 6L6, 807, 5881, and KT66 tubes. The primary impedance is 6,600 ohms plate to plate, and 1,220 ohms screen to screen, to provide the optimum combination of triode and pentode characteristics; the dotted boxes in Fig. 27 show changes from the original Williamson circuit.

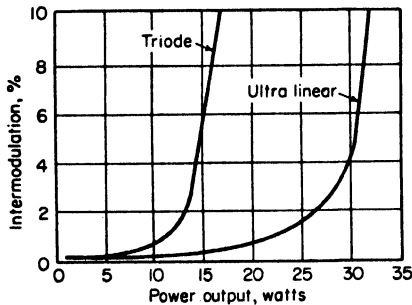



FIG. 28. Intermodulation distortion, 40 and 2,000 cps mixed in 4:1 ratio.

The designers recommend C_9, a 250-μf by-pass across the output-tube cathode resistors. The output power is about 30 watts with 1-volt drive. The ultralinear circuit may be also adapted to push-pull 6550 tubes with 60 watts output.[2]

26. "Circlotron" 50-watt Power Amplifier.[3] A scheme to improve output-transformer characteristics is part of Fig. 29. Two identical d-c power supplies are provided, each supplying half the total power-stage current. The output tubes are in series for d.c., but in parallel for audio signals.[4] Because of the parallel arrangement, only one-fourth

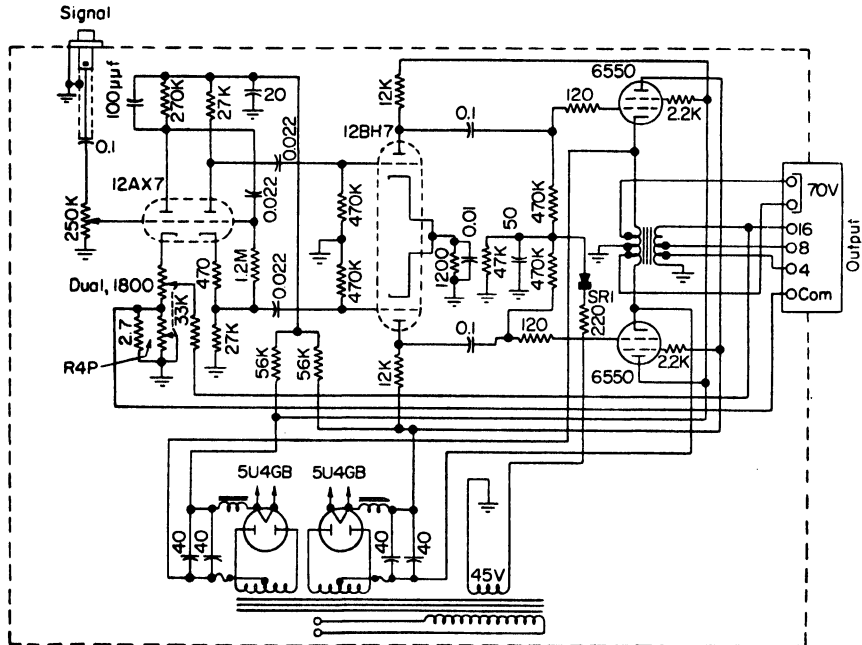



FIG. 29. "Circlotron" 50-watt amplifier, Model A50, with dual power supplies, split load and special output transformer. (*Courtesy of Electrovoice, Inc.*)

[1] Hafler, D., and H. I. Keroes, "Ultra Linear Amplifiers," U.S. Patent 2710312, June 7, 1955. Keroes, H. I., "Audio Transformer," U.S. Patent 2680218, June 1, 1954. D. Hafler and H. I. Keroes, Improving the Williamson Amplifier, *Radio and Television News*, **49**, 43, February, 1953. D. Hafler and H. I. Keroes, An Ultra-linear Amplifier, *Audio Eng.* **35**, 16, November, 1951.

[2] Keroes, H. I., Adapting the Ultra Linear Williamson to 6550 Operation, *Radio and Television News*, **54**, 52, November, 1955.

[3] MacPherson, C. H., Unity Coupled Power Amplifier, *Service*, **44**, January, 1955.

[4] Hall, C. T., "Parallel Opposed Power Amplifier," U.S. Patent 2705265, Mar. 29, 1955.

the usual primary inductance is needed, with a consequent improvement by decreasing leakage inductance and distributed capacitance. With balanced 6550 tubes, no d.c. flows in the output transformer. The total output primary winding presents a load to each of the two output tubes. One-half the load is in the cathode and the other half in the plate circuit, the plate load of one tube being the cathode load of the other (Fig. 30). Because each tube looks at the same load as the other, the result is unity coupling between tubes. Despite the residual leakage reactance in the transformer, no switching transients occur in operation, for both tubes have the same signal current flowing.

FIG. 30. Circlotron circuit simplified.

This amplifier is rated at 50 watts ±0.5 db from 20 to 75,000 cps, less than 0.5 per cent harmonic distortion, and less than 1 per cent intermodulation distortion at rated output; hum and noise are 85 db down. Input required is 1.25 volts rms across a ¼-megohm load.

This circuit offers a variable-output damping factor from 0.1 to 10. For Klipsch speakers, for example, the damping factor is set at 10; for speakers on infinite baffles it is set at 1.0; seldom is it used below 0.5. This is accomplished by a ganged potentiometer. Current feedback to the input 12AX7 is in series with the common output and voltage feedback from the 16-ohm output to R_{4p} in the cathode circuit of the 12AX7.

FIG. 31. McIntosh 60-watt amplifier with unity coupling output transformer.

27. McIntosh 60-watt Amplifier.[1] Figure 31 is the circuit for the McIntosh MC-60 amplifier which uses push-pull 6550 tubes in a patented output circuit. This is a class B

[1] McIntosh, F. H., "Wide-band Amplifier Coupling Circuit," U.S. Patent 2477074, July 26, 1949.

circuit designed to eliminate the "notch" which occurs in the output wave form when a conventional push-pull amplifier operates both class A and class B. The driver is a push-pull cathode follower 12AX7.

The bifilar-wound transformer primary[1] gives extremely tight coupling and a minimum leakage reactance between primary windings. Each tube works into two primary sections, **one in its plate circuit and one in its cathode circuit; these sections are wound with unity coupling. This provides some feedback, but an additional output transformer winding feeds the input 12AX7 cathode by way of R_{11}, C_6, and R_7 to ground.**

Harmonic distortion is 0.3 per cent from 20 to 20,000 cps at any output level up to 60 watts. The input is high-impedance requiring 0.5 volt across 0.5 megohm.

MULTIPLE-SPEAKER SYSTEMS

28. Crossover Networks. For reproduction of the complete audio spectrum, two or more speakers may be required, a large speaker (woofer) for the low end of the spectrum and a small speaker (tweeter) for the high frequencies. In some instances

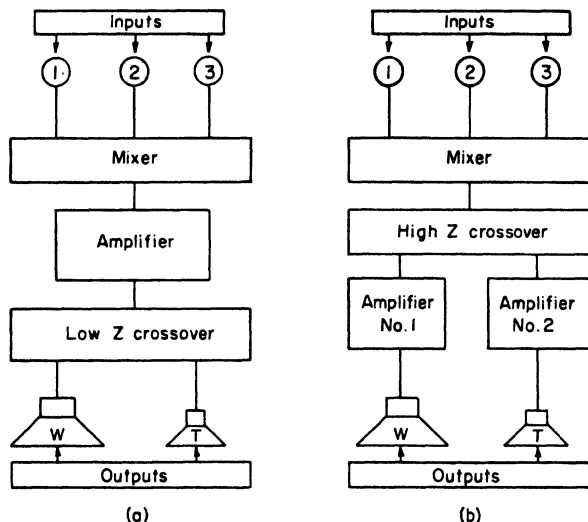



FIG. 32. (*a*) Output crossover; (*b*) input crossover requiring two amplifiers.

a third speaker for the mid-range is employed. A crossover or "dividing" network is provided so that only the part of the spectrum for which a particular speaker is adapted is allowed to reach it.[2]

There are two general types of crossover networks: low-impedance, which are used at the output of a single amplifier (Fig. 32*a*), and high-impedance, which are used at the input of two amplifiers (Fig. 32*b*).

It is the purpose of the crossover network to divide up the ranges of frequencies, and for the low-impedance type to provide a reasonably constant output load Z_0 on the amplifier for all frequencies. Paper capacitors and air-core coils are used for the output types.

29. Two-way Output Networks, 6 Db per Octave. A series circuit is shown in Fig. 33*a* and a parallel circuit in Fig. 33*b*. The values for L and C are found from Fig. 34. Note that the bottom horizontal scale is in octaves from 125 to 4,000 cps. Following straight up from the desired frequency, there is a triple intersection: straight left is the reactance X,

[1] Corderman, Sidney A., and F. H. McIntosh, A New 30-watt Power Amplifier, *J. Audio Eng. Soc.*, **1**, 292, October, 1953.

[2] Hitchcock, R. C., A Survey of Crossover Networks, *Audio*, **40**, 28, November, 1956.

slant up left along dashed line is the capacitance, slant up right along the dotted line is the inductance. For example, at 500 cps the reactance is 16 ohms for a 20-μf capacitor and a 5.1-mh inductor. These are the values required for both circuits of Fig. 33 for a 16-ohm woofer W and a 16-ohm tweeter T, with a crossover at 500 cps, fed from an amplifier output Z_0 of 16 ohms.

30. Two-way Output Crossover Network, 12 Db per Octave. Figure 35, in which two capacitors and two inductors are used, for either the parallel or the series circuit, produces a sharper crossover. This type is safer, as it protects the high-frequency driver units from damage by low-frequency signals. Note that the subscripts are not the same for the two circuits. For the series circuit of Fig. 35*a* both inductors and capacitors have 0.71 times the reactance values shown by Fig. 34. Inductor values are linear; to get 0.71 times the reactance, the L value shown by the chart is multiplied by 0.71. Capacitor values are reciprocals; to get 0.71 times the reactance, the C value from the chart is multiplied by 1.41 for Fig. 35*a*.

FIG. 33. Two-way crossover; 6 db per octave.

For the parallel circuit of Fig. 35*b*, the desired reactance values are 1.41 times those of the chart; the inductor values are multiplied by 1.41, and the capacitor values by 0.71 as compared with the chart values. For the series circuit, 500-cps crossover, 16-ohm speakers, C_2 is 28.2 μf and L_2 is 3.6 mh. For the parallel figure under the same conditions, C_3 is 14.2 μf and L_3 is 7.2 mh.

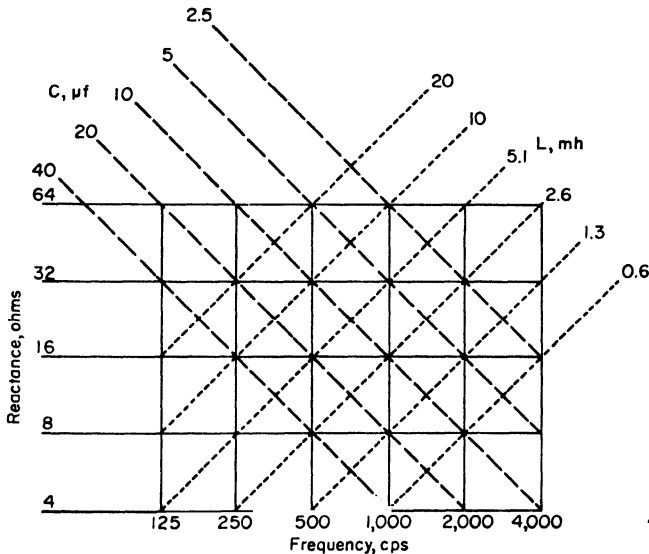



FIG. 34. Reactance chart.

31. Three-unit Output Crossover. Figure 36 is a simple three-unit crossover. For 16-ohm speakers let f_1 be 500 cps where W starts to cut in, and f_2 be 4,000 cps where T starts to cut in. Thus L_5 for the woofer has 16 ohms at 500 cps, 5.1 mh from the chart. For the tweeter, C_5 has 16 ohms at 4,000 cps, 2.5 μf from the chart. The mid-range speaker M has a double-capacitance capacitor (half reactance) and a half-inductance coil (half reactance) in series with it. Thus C_4 has 8 ohms at 500 cps, 40 μf, and L_4 has 8 ohms at 500 cps, 2.5 mh.

A typical commercial three-way output crossover network[1] is shown in Fig. 37. As shown the amplifier has 16-ohm output, the three speakers are 16 ohms each, and the crossover frequencies are 500 and 5,000 cycles.

32. Two-way Input Network.[2] Figure 38 is a double back-to-back *RC* network designed to furnish two separate power amplifiers with high and low frequencies, respectively, each amplifier having its own loud-speaker. A cathode follower ahead of the network is recommended. Constants shown are for a crossover at 300 cycles and $R = 50{,}000$ ohms. At

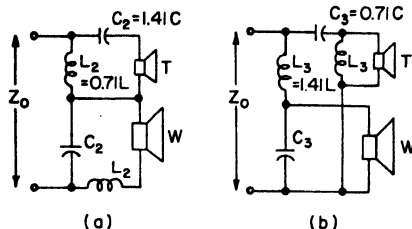



FIG. 35. Two-way crossovers, 12 db per octave.

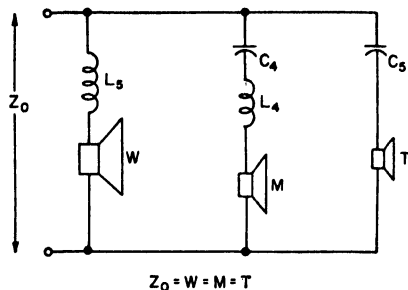



FIG. 36. Three-way crossover. For low-frequency crossover f_1, find C and L from Fig. 34. $L_5 = L$; $L_4 = L/2$; $C_4 = 2C$.

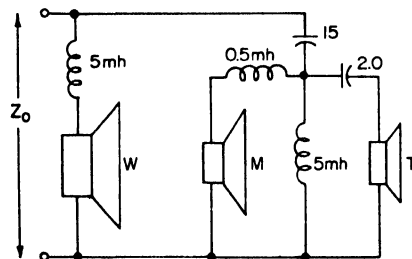



FIG. 37. Three-way crossover from Klipsch; $Z_o = W = M = T = 16$ ohms.

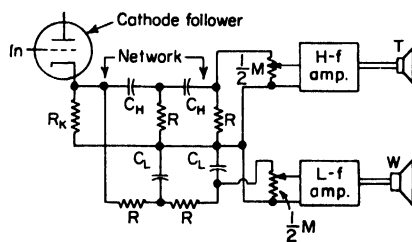



FIG. 38. Two-way crossover, high impedance; for 300 cps crossover, $R = 50$ kilohms; $C_H = 0.02$ μf; $C_L = 0.005$ μf.

twice the crossover frequency, $f_L = 2f_c$, the network output voltage is 12 db down (one-fourth voltage) for the low-frequency filter, and at half the crossover frequency, $f_H = f_c/2$, the voltage is 12 db down for the high-frequency filter. With subscripts L for low, H for high, and f_c for the crossover frequency:

$$f_c = f_L/2 = 2f_H \text{ cycles}$$
$$C_L = 1/(2\pi f_L R) \text{ farads} = 159{,}000/(f_L R)\ \mu\text{f}$$
$$= 4C_H$$
$$C_H = 159{,}000/(f_H R)\ \mu\text{f}$$

33. Two-way Input Crossover with No Insertion Loss. The simple circuit of Fig. 38 has no amplification since it has no tubes. Its output is one-fourth the input or is 12 db down. This may be handled by some amplifier systems but in other cases an additional stage of amplification will be needed. A versatile electronic crossover using 12AX7 tubes having no insertion loss is shown in Fig. 39. Each channel is adjustable; various responses are shown in Fig. 40. The normal operating levels are 2.0 volts input and output with a maximum of 5 volts. The low-frequency channel is flat within ±1.2 db from 10 cps to cutoff and the high-frequency channel is ±12.3 db from cutoff frequency to 20 kc.

[1] Klipsch, Paul W., private communication.
[2] Stone, Lewis C., At Home with Audio, Hi-fi Segregation, *Audio*, **40**, 24, May, 1956.

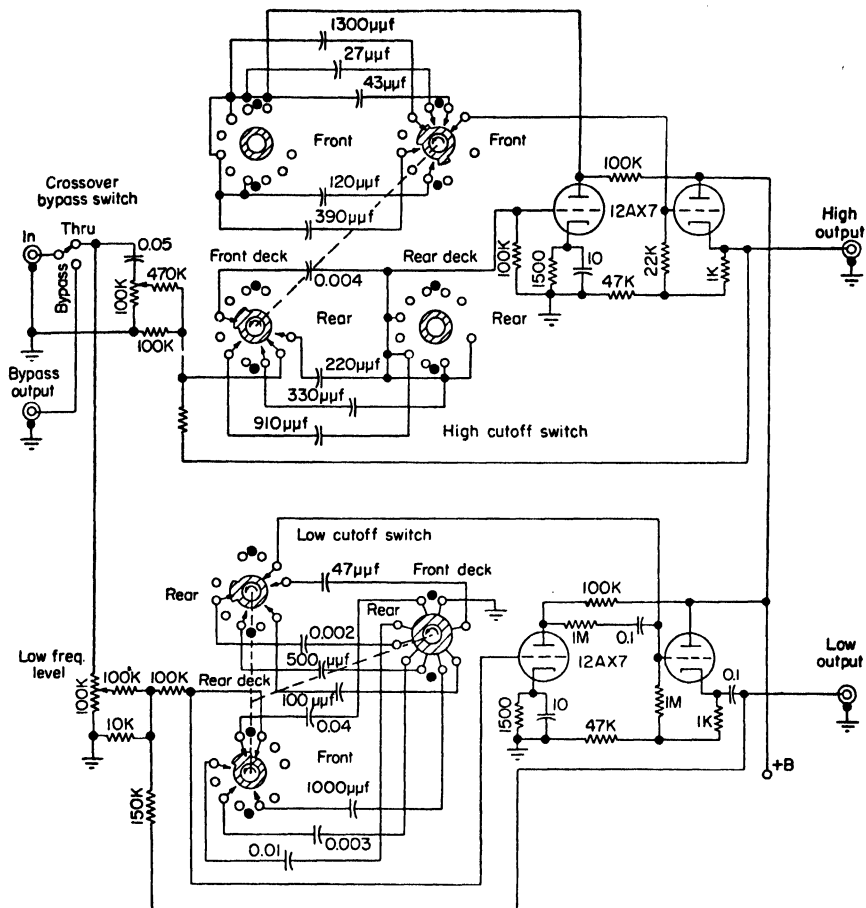



FIG. 39. Heathkit XO-1 electronic crossover having no insertion loss.

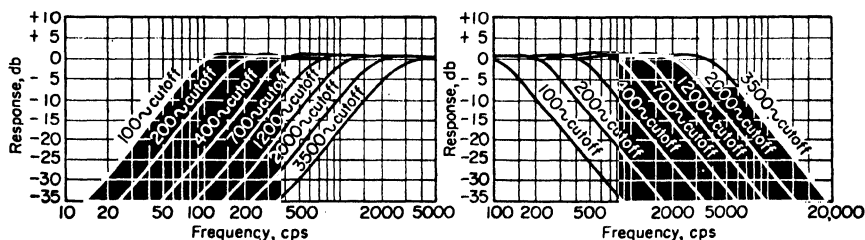



FIG. 40. Characteristics of XO-1 crossover.

34. Multiple-speaker Power Relations. In some installations, a large number of identical permanent-magnet speakers are operated from a single amplifier output. The speakers can be in series where each has the same audio current, in parallel where each uses the same output voltage, and combinations of these two. The

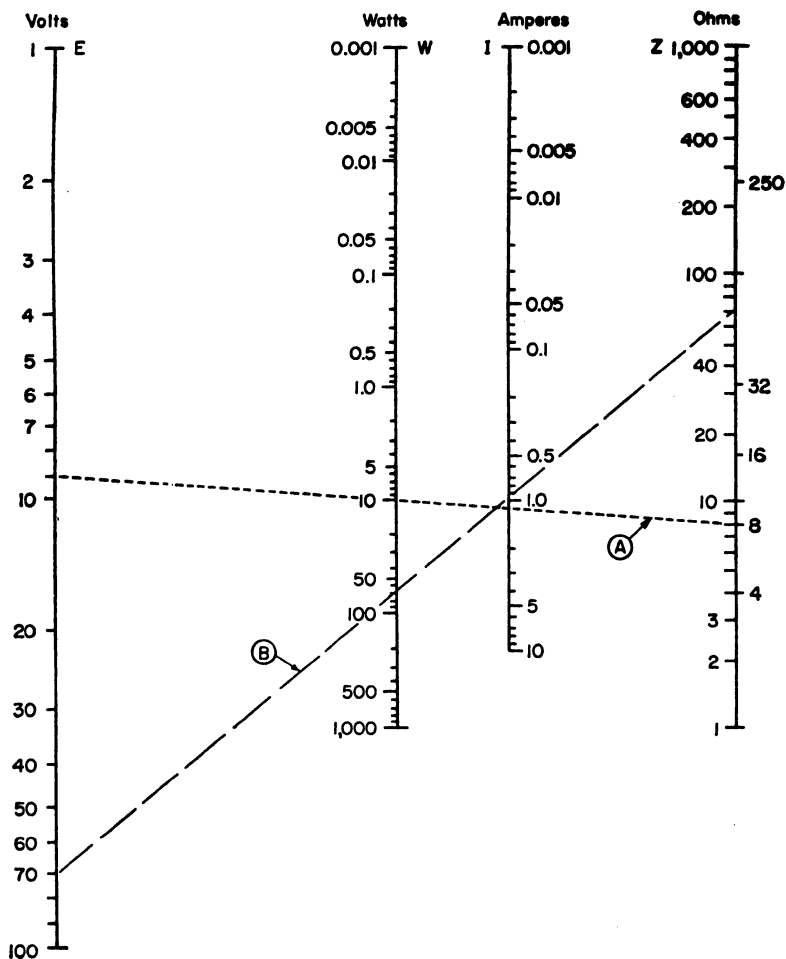



FIG. 41. Ohm's law chart. (*A*) 8-ohm speaker; (*B*) 70-watt amplifier.

amplifier rated output is P_0 watts, E_0 volts, I_0 amp, and Z_0 ohms. For any individual speaker in the installation, the relations are

$$P_i = (E_i)^2/Z_i \qquad \text{and} \qquad I_i = E_i/Z_i$$

where the subscripts i indicate individual loud-speakers. Figure 41 shows these relations in chart form.[1] The dotted line on the chart shows a single 8-ohm speaker rated at 10 watts to require 9.0 volts and 1.25 amp.

[1] Hitchcock, R. C., Ohm's Law Charts, *Elec. J.*, October, 1933, p. 435.

When n speakers are used in parallel, $m = 1$; when m speakers are used in series, $n = 1$; for series-parallel, n is the number in parallel and m the number in series (Fig. 42).

FIG. 42. Multiple-speaker arrangements.

Several combinations of standard 4/8/16-ohm speakers can be made by a judicious choice of the distribution circuit. As an example, consider a 70-watt amplifier with 4/8/16/72/143-ohm outputs. If the parallel circuit is used with n speakers, each must have nZ_0 impedance. Using the 4-ohm output connection, $n = 1$ would take one 4-ohm speaker; $n = 2$, two 8-ohm speakers; $n = 4$, four 16-ohm speakers. For the 8-ohm connection, $n = 1$ would take one 8-ohm speaker and $n = 2$ two 16-ohm speakers.

The series circuit has more possibilities: $m = 1$ would be one 4-ohm speaker on the 4-ohm tap; $m = 2$, two 4-ohm speakers in series on the 8-ohm tap; $m = 4$ allows four 4-ohm speakers on 16-ohm tap, and on up to 36 series 4-ohm speakers when $m = 36$ on the 143-ohm tap. The series-parallel circuit has a great many possibilities. These can be worked out from Table 3, and by reference to Fig. 42.

Table 3. Series and Parallel Loud-speakers

No. of speakers in parallel	No. of speakers in series	P_i, watts	E_i, volts	I_i, amp	Z_i, ohms
n	1	P_0/n	E_0	I_0/n	nZ_0
1	m	P_0/m	E_0/m	I_0	Z_0/m
n	m	P_0/mn	E_0/m	I_0/n	$(n/m)Z_0$

Normally a power amplifier has enough feedback to maintain the output voltage practically independent of the load. Thus a tap on the output transformer can be regarded as being a fixed voltage line. Available on some output transformers is a tap specifically marked "70.7" volts or "100 volts." If the transformer connections are marked only in ohms, the proper tap for a fixed voltage can be found by referring to Fig. 41. For instance, a 70-watt amplifier may have 71 volts output at the 72-ohm tap (dashed line on Fig. 41), with an output rated current of 1.00 amp. Each transformer, primary has $Z_i = (E_i)^2/P_i$, where $E_i^2 = 5{,}000$ ohms for the 70.7-volt line and 10,000 ohms for the 100-volt line. Each loud-speaker takes P_i primary watts. As an example, a 70-watt amplifier with a 70.7-volt output connection may have four 10-watt speakers connected, each of whose transformer primaries has 5,000/10 = 500 ohms, plus five 6-watt speakers each of whose transformer primaries has 5,000/6 = 833 ohms. The total watts is $4 \times 10 + 5 \times 6 = 70$ watts. Of

course, fewer speakers can be used and less output employed. Each speaker transformer is matched to its individual speaker voice coil. That is, if the 6-watt speakers are 4-ohm units, each transformer is 833 ohms to 4 ohms.

Constant-line-voltage transformers are also available which have taps to match standard voice-coil impedances for specified watt ratings. The loss in the matching transformers is usually neglected, because an 80 per cent efficient transformer is only 1 db down in power at its secondary.

35. Phasing Multiple Speakers. When many speakers are used and where two or more may be heard at one spot, such as a ball field or a gymnasium, it is necessary

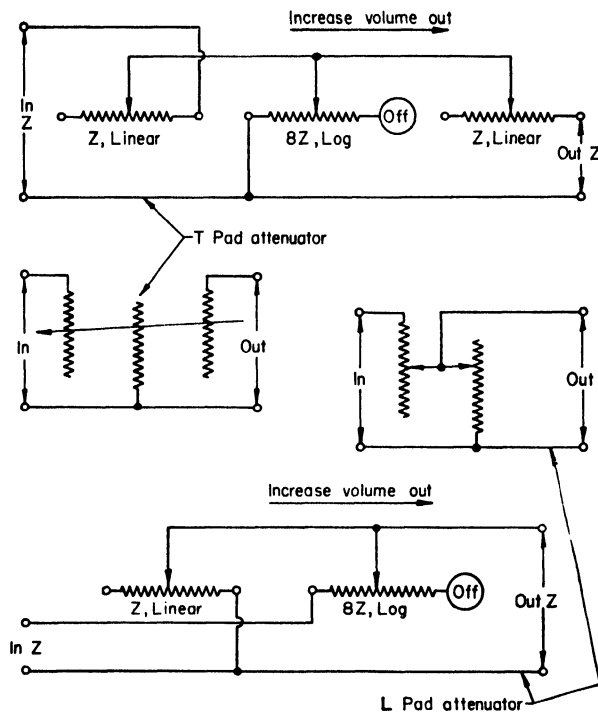



FIG. 43. Variable pads to control speaker volume.

to have each speaker correctly phased. This is readily checked with a dry cell by observing each cone motion as the voltage is applied. Each voice coil should move out when the others move out. If one moves the wrong way, its terminals should be reversed. In general, speakers are made alike and a little care in symmetrical wiring will ensure correct phasing.

36. Individual Volume Controls. In some cases it is necessary for each speaker to have its own volume control, as in hotels, hospitals, etc. An inexpensive L pad or T pad may be used. The L pad has two ganged potentiometers; the T pad has three. They are wired as indicated in Fig. 43. Control from practically zero to full volume does not materially affect either the line or the other speakers. Note that, in each pad, one of the controls has a logarithmic taper and an "off" position at maximum volume output. These pads are available for 4/8/16 ohms, and as high as 3,000 ohms. Their rating is 4 watts continuous d.c. or 15 watts peak.

37. Stereophonic Reproduction. A typical system of single-channel recording or reproducing is shown in Fig. 44*a*; a distributed source of sound is picked up by three input microphones M_1, M_2, M_3, combined to give a single composite signal, fed to an amplifier, then to the three identical speakers. A sound primarily at the left will be supplied equally to each of the three speakers.

In Fig. 44*b*, the same distributed sound source energizes the three microphones spaced as in the previous figure. But now each microphone feeds its own amplifier, there is no mixing, and each amplifier feeds its own speaker. A signal which is louder at microphone M_1 comes out louder in speaker S_1 than in the others. This is stereophonic sound—two or more pickups and output reproducers which give "three"-dimensional sound.[1]

Good results can be obtained by two independent channels. The widest possible spacing of the speakers on the long dimensions of a rectangular room is apparently better than when arranged on the narrow end of such a room.[2] Two-channel stereophonic

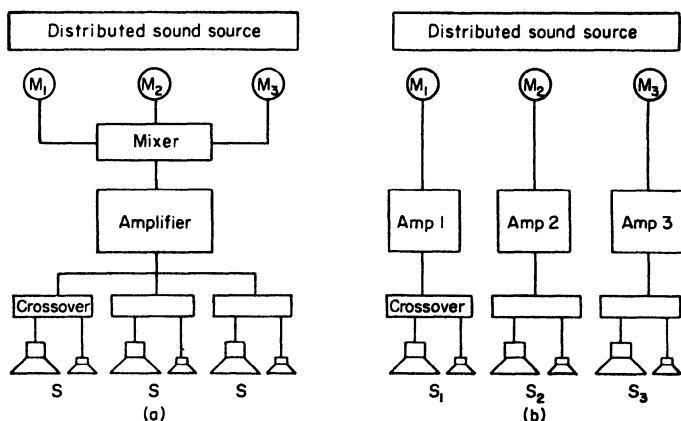



FIG. 44. (*a*) Single channel, three identical outputs; (*b*) three channels, stereophonic output.

phonograph records, requiring two identical pickup heads, and stereo tapes which require dual pickup heads are available. Two complete amplifiers and loud-speaker systems are needed; the realism is excellent.

COMPONENTS

38. Output Transformer. When an output transformer is used it is the most critical component in an audio amplifier. See also Sec. 49. The three most important transformer constants are the turns ratio, the primary inductance, and the leakage inductance.[3] For high audio quality it is also important to have low harmonic and intermodulation distortion. These are obtained by using a low magnetic flux, a good grade of transformer iron, and low d-c copper resistance.

The turns ratio is given by $T_{pri}/T_{sec} = \sqrt{Z_{pri}/Z_{sec}}$, where the Ts are the turns, and the Zs the impedances. For triodes the Z_{pri} required is twice to four times the plate resistance, the higher value giving lower distortion. For example for a 10,000-ohm primary and a 16-ohm secondary, the turns ratio is $\sqrt{10{,}000/16} = 25:1$. This will be suitable from plate to plate of triodes of $\frac{1}{2} \times 10{,}000$ to $\frac{1}{4} \times 10{,}000 = 5{,}000$ to 2,500 ohms.

[1] Tinkham, R. J., Binaural or Stereophonic? *Audio Eng.*, **37**, 22, January, 1953. Denny, Wayne B., A Dual Channel Control Amplifier for Stereophonic Music Systems, *Audio*, **38**, 17, May, 1954.

[2] Klipsch, Paul W., Experiences in Stereophony, *Audio*, **39**, 16, July, 1953.

[3] Hitchcock, R. C., and W. O. Osbon, Design of the Output Transformer, *Electronics*, November, 1930, 381; December, 1930, p. 427.

The inductance required to obtain 10,000 ohms depends on the low-frequency response expected. Thus, if f = 20 cycles,

$$L = Z_{pri}/2\pi f = 10{,}000/(2\pi 20) = 80 \text{ henrys}$$

This is readily measured by 60-cycle power, with a voltmeter and an ammeter, where $2\pi fL = Z = E/I$; $L = E/(2\pi fI)$ with the secondary open-circuited.

The leakage inductance determines the high-frequency response.[1] To measure leakage inductance, short-circuit the secondary and measure the corresponding primary inductance at a frequency so low that the distributed capacitance is not effective, *e.g.*, 60 cycles.

The response will be 3 db down from the middle-frequency output when $\omega L_s = R_B$, where L_s is the leakage inductance and R_B is the sum of the tube plate resistance and the load resistance. Thus to be down no more than 3 db at 60,000 cps for a load of 10,000 ohms (referred to the primary) and a tube resistance of 2,500 ohms

$$L_s = 12{,}500 \div 376{,}000 \text{ or } 0.033 \text{ henry or } 33 \text{ mh}$$

When a single output tube is used, the inductance of the output transformer must be measured with the appropriate d-c flowing.

A complete design of an excellent 15-watt output transformer is given in the Williamson series.[2] The construction details refer to British core iron, wire sizes, and insulation, but the over-all performance can be matched by a first-rate transformer designer. The original Williamson output transformer has 10 primary sections interleaved with 8 identical secondary sections. Primary sections are permanently connected in series aiding, and only three leads are brought out for the 10,000-ohm plate-to-plate impedance and center tap. All secondary leads are brought out. To give the desired coupling, all secondaries are used at all times, from 1.7 ohms (all in parallel) to 109 ohms (all in series) in the original design. The essential characteristics are primary inductance 100 henrys measured at 50 cycles at 5 volts rms, primary d-c resistance 250 ohms, leakage inductance 22 mh.

Transformer connections can be determined by use of simple equipment in case the designations of the various windings have been lost. The procedure as applied to an output transformer is given below.

With an ohmmeter find out how the several windings are related to each other, *i.e.*, if they have any electrical connection, and to determine which the several windings may have been designed for. Apply a-c voltage, 60 cps, from a Variac to the winding of the highest d-c resistance (assuming that this is an output transformer). Measure the a-c voltage across each pair of leads. Calculate the turns ratio from the ratio of primary voltage to the voltages measured across the several pairs of leads. Calculate the impedance ratio (turns ratio squared) for each set of values.

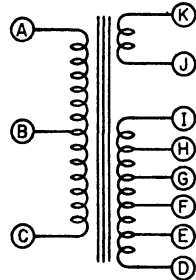



Fig. 45. To determine transformer data.

With the secondaries open, measure the a-c current through the primary with known voltage applied. This gives the impedance of the primary. Assume that this value is four times the plate-to-plate impedance of the tubes to which it will be connected. Finally, using this impedance value, calculate the impedances the several secondary windings are designed for.

Example: The transformer has three windings, one with a center tap, one with numerous taps, and one with no tap. In Fig. 45 is the relation among the windings as determined by d-c resistances and in Table 4 are the various measured and calculated values. In this case it was found that, with 50 a-c volts, 60 cps, applied to the primary a current of 3.3 ma resulted, indicating a reactance of 15,200 ohms. This accounts for the value 3,800 in the table. Winding J to K is a feedback winding.

39. Output Tubes. Tables 5 and 6 are compiled from tube-data sheets. Beam-power tubes are variously called "tetrodes," "beam-power tubes," "power pentodes,"

[1] Scroggie, M. G., "Radio Laboratory Handbook," 4th ed., p. 277, Iliffe and Sons, Ltd., London, 1948; Lee, Reuben, "Electronic Transformers and Circuits," 2d ed., Chap. 5, Wiley, 1955.
[2] Williamson, D. T. N., *Wireless World*, **55**, 282, 1949.

Table 4. Transformer-characteristic Determination

Leads	D-c resistance, ohms	A-c volts	Turns ratio	Impedance ratio T^2	Impedance calculated	Load impedance
A to *B*	90					
B to *C*	80					
A to *C*	170	125	.....	.....	3,800	3,600-4,600
D to *E*	0-1	4.0	31.3	980	3.9	3-4
D to *F*	0-1	5.4	23.1	535	7.1	6-8
D to *G*	0-1	7.2	17.3	300	12.6	12-16
D to *H*	15	29.5	4.24	17.9	212	200-250
D to *I*	21	41.6	3.0	9.0	423	400-500
J to *K*	14	12.5	10	100	38	Feedback

Table 5. Beam-power Output Tubes, Pentode-connected, Two Tubes AB-1 Push-pull

Tube	Output, watts	% distortion	Load, ohms	E_p	E_{g2}	E_{g1}
50L6	8	10	8,000	200	125	−8
6AQ5	10	5	10,000	250	250	−15
6V6GT	14	3.5	8,000	285	285	−19
6L6G	24.5	4	9,000	360	270	−21
5881	26.5	2	6,600	360	270	−23
1614	26.5	2	6,600	360	270	−23
807	36	NA	6,800	400	300	−30
KT66	50	5	6,000	480	385	−27
6CA7	35	5	3,400	375	375	−25
6550	41	4	4,500	400	300	−23
6146	70	NA	5,500	500	185	−40

NA = no data available.

or just "pentodes." Each has a beam-forming electrode connected internally to the cathode. In a voltage pentode this would be the suppressor.

As an approximation, the maximum audio-power output available from two identical push-pull beam pentode tubes, class AB-1, is equal to half the rated dissipation of the plates and grids of the two tubes. If a low figure of distortion is important, the output power will be less than this. Also, to an approximation, the maximum audio-output power from identical push-pull triodes, class AB-1, is about one-fourth the plate dissipation of the two tubes.

Table 6. Triode Output Tubes, Two Tubes, Class AB-1, Push-pull

Push-pull triodes	Plate diss. both tubes d-c watts	Output watts a.f.	Ratio a.f./d.c.	Total harmonic distortion, %
2A3	30	10*	0.33	2.5
		15†	0.50	5.0
6BL7	12	3.5	0.29	NA
6BX7	12	3.5	0.29	NA
6L6	38	5	0.13	NA
KT66	50	14.5	0.29	3.5
6550	80	28	0.35	2.5
6AS7G	26	6	0.23	NA
6080	26	6	0.23	NA
6336	60	15.5	0.26	NA
6337	80	25	0.31	NA

NA = figure not available.
* Cathode bias.
† Fixed bias.

Push-pull vs. Single Tube. More than twice the output can be obtained from two push-pull tubes than for one such tube alone. Usually this figure is 2.2 for beam pentodes, because of the allowable increased bias. The converse is also true; a single class A tube can handle 0.45 times the output power of two such tubes in push-pull.

Table 7. Preamplifier and Interstage Tubes

Tube	Heater		Plate max volts	Triode voltage amplification factor	Triode-connected pentode or triode		Pentode-connected pentode or triode	
	Volts	Amp			*Gm*, μmhos	Plate resistance, ohms	*Gm*, μmhos	Plate resistance, megohms
6SC7[b,d,f]	6.3	0.3	250	70	1,325	53,000		
6SL7GT[b,d]	6.3	0.3	300	70	1,600	44,000		
12SL7GT	12.6	0.15						
12AX7[c,d]	6.3	0.3	300	70	1,600	44,000		
	12.6	0.15						
12AY7[c,d]	6.3	0.3	250	44	1,750	25,000		
	12.6	0.15						
5879[c,e]	6.3	0.15	250	21	1,530	13,700	1,000	2.0
12AT7[c,d]	6.3	0.3	300	60	5,500	10,900		
	12.6	0.15						
6N7[b,d,f]	6.3	0.8	300	35	3,200	8,000		
6BK8/Z729[c,e]	6.3	0.2	300	..	.....	......	1,850	2.0
6SN7GTA[b,c]	6.3	0.6	450	20	2,600	7,700		
12SN7GT[b,c]	12.6	0.3	300	20	3,000	6,700		
6AU6[a,e]	6.3	0.3	250	36	4,800	7,500	5,200	1.0
12AU6[a,e]	12.6	0.15						
12AU7[c,d]	6.3	0.3	300	20	3,000	6,700		
	12.6	0.15						
6BZ7[c,d]	6.3	0.4	250	38	6,800	5,600		
12BH7[c,d]	6.3	0.3	300	16.5	3,100	5,300		
	12.6	0.15						

[a] (7) miniature 7-pin base.
[b] (8) octal base.
[c] (9) miniature 9-pin base.
[d] (dt) dual triode.
[e] (p) pentode.
[f] (s) single cathode.

Triode Output Tubes. The class A output for triode-connected pentodes with grid No. 2 connected to the plate is about one-third that of the same tubes class AB-1 as pentodes. For classes A and AB-1 no control-grid current may be drawn; so the driving peak voltage must be less than the bias voltage. With −15 volts grid No. 1 bias, a single tube may be driven to a value just less than 15 volts. Since the rms value of a sine wave is 0.707 times its peak, this fixes the allowable rms drive voltage. Thus −15 volts peak has $0.707 \times (15) = 10.6$ volts rms. A pair of push-pull tubes, each biased at $-E$ volts, can be driven with $2E$ peak volts grid to grid, or $1.41E$ volts rms grid to grid. This applies to either pentode or triode classes A and AB-1.

40. Interstage and Input Tubes. The major criteria for a good voltage amplifier tube are voltage gain, freedom from noise (leakage, thermal, and microphonics), and minimum hum due to a-c heater. If one or more of the above are neglected, there are literally dozens of tubes which would be suitable for audio amplifiers. There are at least three tubes which have been designed specifically for audio work, meeting the above criteria. These are 5879,[1] 6BK8/Z729,[2] and 12AY7. The minimum hum for any dual heater tube is obtained with the 6.3-volt connection. When these tubes are used on d.c. the 12.6-volt connection may be used. Table 7 includes the special audio tubes and some others often used. Table 8 shows some typical operating conditions for 5879 and 12AY7.

[1] Heacock, D. P., and R. A. Wissolik, "Low-noise Miniature Pentode for Audio Amplifier Service," Publication ST-598, Tube Dept. RCA, 1951. Reprinted from *Tele-Tech*, February, 1951.
[2] Applications Laboratory Report 17, 6BK8/Z729, *Radiotronics*, **20**, 61, May, 1955.

Table 8. Typical Low-noise Audio-tube Operating Conditions

	5879 pentode		12AY7 dual triode, each unit		
	Max voltage output	Max voltage gain	Low-impedance drive, approx 200 ohms	High-impedance drive, approx 100 kilohms	Low-level service
Plate supply, volts	300	300	300	300	150
Plate load, megohms	0.47	0.47	0.51	0.51	0.020
Grid No. 2 resistor megohms	1.2	1.9			
Grid No. 1, next stage, megohms	1.0	1.0	1.0	1.0	1.0
Cathode resistor, ohms	2,500	1,200	6,000	6,700	2,700
Grid No. 2 by-pass μf	0.1	0.1			
Cathode by-pass, μf	4.3	4.3	2.0	2.0	40
Plate coupling capacitor, μf	0.004	0.004	0.004	0.004	0.004
Peak output, volts	110*	42*	64*¶	76*¶	
Voltage gain	76†	152‡	31	30	12.5

* On grid No. 1 resistor of next stage.
† At an output of 5 volts rms.
‡ At output of 1 volt rms and grid No. 1 bias of 1 volt.
¶ This is peak (1.41 times rms) for approximately 5 % total harmonic distortion.

41. Resistors and Capacitors. Audio amplifiers use several types of resistors and capacitors. Fixed resistors are usually of the carbon-filament type provided with tightly fitted external insulating shells. The units purchased to standard ±5 per cent tolerances are used in matched push-pull circuits, for example, R_1 and R_2 in Fig. 46. In the 1-megohm range these resistors will hold their values within ±5 per cent at 40°C ambient and with a 40°C rise in temperature.

R_1 and R_2 can be checked for balance by reading d-c voltages E_1 and E_2, which, if V_1 and V_2 are matched, will read the degree of static balance. Further checking can be made if V_1 and V_2 are separate tubes, *i.e.*, not a dual-type tube, by interchanging them and again reading E_1 and E_2. If the voltages come out the same as before, I_1 and I_2 are alike and equal to E_1/R_1 and E_2/R_2.

Fig. 46. D-c checks of push-pull stages.

Precision carbon resistors with 1 per cent tolerance are also available in ½-, 1-, and 2-watt ratings. These are slightly larger than the

others for the same rating. Precision resistors usually have low noise and are desirable for preamplifier stages. The temperature coefficient of resistance of one type varies from −0.005 per cent for 10-ohm resistors to −0.02 per cent per degree centigrade rise for the ½-megohm resistors—about half that of ordinary resistors.[1] These precision resistors are available either insulated or noninsulated. A noninsulated resistor runs cooler per watt than an insulated resistor.

Vulcanized resistors are also made which have very low noise, a typical specification "corresponding to a change of resistance of one part in a million over the audio range."[2] These resistors have a temperature coefficient varying from +0.2 per cent for 5,000-ohm units to −0.4 per cent per degree centigrade for 10-megohm resistors, in the temperature range 27 to 77°C.

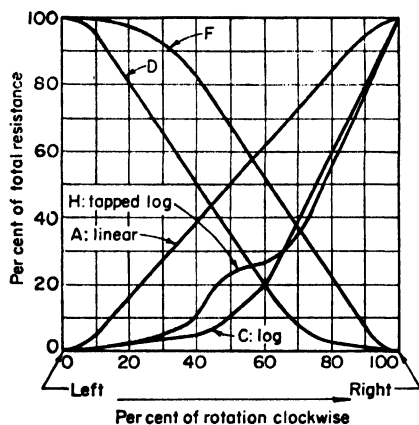



Fig. 47. Potentiometer tapers.

Fixed resistors with low ohmage values are wire-wound (WW) and available in 4, 5, 7, 10 watts and higher. A typical temperature rise is 290°C for full rated load on a 10-watt wire-wound resistor.[1] In Fig. 46 R_9 would probably be wire-wound to handle the output-tube plate currents.

Temperature-sensitive resistors are desirable in some power-supply systems; see Sec. 42 and Table 9.

Potentiometers are used for volume and tone controls and hum adjustment. Wire-wound controls have linear curves but carbon or composition types may be any one of the characteristics shown in Fig. 47. In logarithmic curve C tapers, 10 per cent of the resistance is contained in a 50 per cent rotation of the arm. This is "audio" taper.

For audio amplifiers mica capacitors are recommended where high power or high frequencies are encountered as in feedback circuits; see C in Fig. 76.

Coupling capacitors such as C_4 and C_5 of Fig. 46 are paper types. For the output stage to receive a balanced signal at all frequencies, the C_4, R_6 circuit must match the C_5, R_8 circuit. Thus both capacitors and resistors must be matched. The capacitors must also have high insulation resistance. To test this, with tubes V_3 and V_4 removed, the d-c currents I_4 and I_5 should be zero. When V_3 and V_4 are inserted, grid currents I_7 and I_8 should be very low, at most a few microamperes for class AB-1. If C_4, for example, allows an appreciable current to flow, the bias of V_3 will be wrong, as current will flow through R_1 and C_4. Paper capacitors are also used for by-pass. Here, too, the leakage must be very low, to ensure proper voltage applied to grid No. 2 of V_1.

Electrolytic capacitors may have an insulated outer shell of cardboard, or an outer conducting shell of aluminum. The shell is the negative terminal. The polarity must be observed.

In Fig. 46 an electrolytic C_9 by-passes R_9, the output-tube common-cathode resistor. Leakage current of a fraction of a milliampere per microfarad will be common but I_{10} must be small compared with I_9 for the capacitor to be effective. The reactance of C_9 at the lowest frequency desired must be smaller than the resistance of R_9 in ohms.

POWER SUPPLIES

Power-supply circuits furnish plate, grid, and filament or heater power to the amplifier tubes. The d-c supply for plates and grids includes a rectifier and filter.

[1] International Resistance Co.
[2] S. S. White Industrial Div.

The heater supply is usually a.c. from a 6.3-volt transformer secondary, though series strings of tubes with the same a-c heater current are becoming more popular, now that controlled warm-up-time characteristics are available. For some front-end circuits of high-gain amplifiers, d.c. is supplied to the heaters (see Fig. 25).

42. Power-supply Considerations. Rectifiers are chosen for one or more of the following characteristics: average and peak current, reverse voltage peak, ambient temperature and allowable rise, warm-up time, heater-to-cathode voltage, regulation (no-load to full-load voltage drop), back current (leakage or reverse current).

The power-output tubes take the largest amounts of d.c. and usually are indirectly heated (cathodes electrically but not thermally separated from the heaters). Until these heater-type tubes warm up, the initial d-c voltages are considerably higher than the final or working voltages. Two schemes are available for minimizing this. First, an indirectly heated cathode tube such as 5V4G or 6X5 may be used; or second, a series resistor having a negative temperature coefficient of resistance may be used. The latter have a high resistance when cold and when small currents are drawn, and low resistance when hot and when the normal load current is drawn.

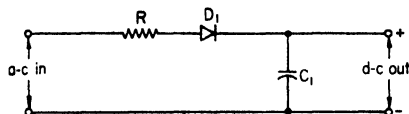



Fig. 48. Half-wave rectifier. R is a current-limiting resistor.

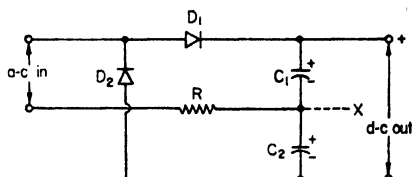



Fig. 49. Full-wave doubler. Point X is used when balanced-to-ground voltages are required.

In Fig. 24, note that a 33-ohm resistor is in series with the positive lead of the rectifier tube. This minimizes the surge of starting current in charging the first filter capacitor. Where the current required is not too high or where the regulation is not important, resistors like the series elements in the filter work satisfactorily as in this circuit. Chokes have less d-c voltage drop and produce a better regulation characteristic than resistors as compensation for their greater weight, volume, and cost. For heavy-duty power amplifiers it is essential that steady d-c voltages be supplied to the screens of the power tubes. Voltage-regulator tubes of the VR(OC) type are employed for this purpose.

Table 9. Globar Thermistors*

Type	Hot† resistance, ohms	Rated load, ma	Working ambient, °C	Cold 25°C resistance, ohms	Rated load, watts, I^2R
439F	200 ± 10%	150	40	1,400 ± 20%	4.5
525F-1	20 + 5% − 15%	600	60	250 + 10% − 30%	7.2
327F-1	35 + 5% − 15%	600	60	460 ± 20%	12.6

* Courtesy of H. R. Emes, Globar Div., Carborundum Co.
† Body temperature 135 to 185°C depending on ambient and proximity of other hot components.

43. Notes on Power Supplies. Figure 48 is a simple half-wave rectifier in which D_1 may be a single diode or half of a double-diode tube or a metal rectifier consisting of several identical units in series if selenium or a single unit if silicon. Note that one d-c terminal is electrically connected to one a-c terminal. This is often advantageous if the common terminal can be grounded but is dangerous if the a-c plug is reversed. In Fig. 49 neither d-c terminal is directly connected to either a-c terminal.

Resistor R prevents excessive surge currents from flowing into the capacitors when first connecting the rectifier to the a-c line. Point X may be used when a balanced-to-ground power supply is needed. This circuit would then amount to two half-wave rectifiers like Fig. 48 connected back to back.

In Fig. 50 one side of the output is connected to one side of the a-c line.[1] In Fig. 51 this is avoided and in addition the capacitor negative terminals can be common.[2]

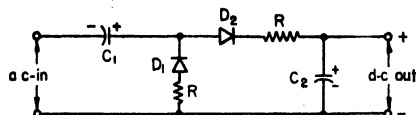



FIG. 50. Half-wave doubler; a-c line common to d-c negative.

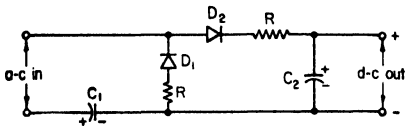



FIG. 51. Half-wave doubler, capacitor negatives common.

Regulation of Half-wave, Full-wave, and Doubler Circuits. Figure 52 shows the comparative regulation for the circuits just mentioned.[3] The input for each circuit is 117 volts 60 cycles. Rectifiers are four 25Z6 tubes, a total of eight diodes. The total capacitance is 48 μf (paper type) $\pm 5\%$. For doubler circuits each "diode" is four parallel diodes, and each capacitor 24 μf.

For all loads the full-wave doubler supplies the greatest voltage. The half-wave doubler falls off rapidly with load, and the half-wave rectifier shows the least change of voltage with increasing load current, *i.e.*, the best regulation. At heavy output current, the voltage on one capacitor of the doubler reverses, and an electrolytic is not recommended.

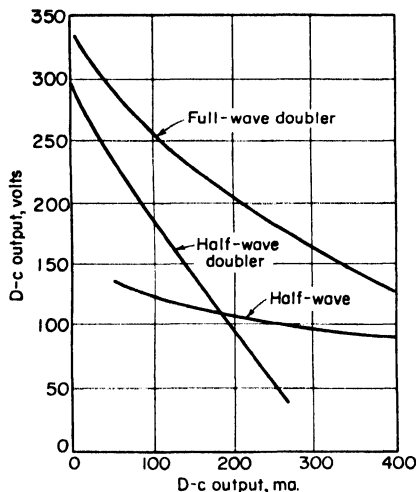



FIG. 52. Rectifier regulation characteristics.

SPECIAL CIRCUITS

44. Phase Inverters. Most audio inputs are single-ended, having two terminals, one at ground potential and one varying according to the audio signal impressed. This "unbalanced-to-ground" system is often carried through the amplifier to the output stage. Here push-pull operation is usually desired. This requires a three-terminal connection, the center being the reference ground and the other two being balanced to ground. The nonground terminals alternate in voltage above and below ground. The various methods for changing unbalanced to balanced circuits are called phase splitters, phase inverters, concertina, paraphase, etc.[4] Criteria in selecting the most suitable phase inverter are linearity, voltage gain, and maximum voltage available.

Split-load Inverter. In Fig. 53 a triode drives a phase-splitting triode. The plate-load resistor is split, one-half being in the plate circuit and the other half being in the cathode circuit R_3 and R_4. Equal resistors R_5 and R_7 are the grid returns for the following stage. The gain of V_2 in Fig. 53 is less than unity; the peak-to-peak voltage from points 1 and 2 is

[1] McIlvaine, O. T., "Rectifier Tube," U.S. Patent 1946354, Feb. 6, 1934.
[2] Montgomery, J. L., "Voltage Doubler Circuit," U.S. Patent 2172962, Sept. 12, 1939.
[3] Hitchcock, R. C., Thermionic Rectifier Circuits, *Electronics*, February, 1944, Fig. 6, p. 102.
[4] Langford-Smith, F., "Radio Designer's Handbook," 4th ed., p. 522, RCA, 1953.

less than that supplied by the driver. Linearity is best for small voltage swings and for matched tubes and resistors R_3 and R_4.

A simplified phase splitter is employed in the Williamson circuit (Fig. 26), where the triode input-tube plate is connected directly to the grid of the phase splitter 6SN7. The voltage drops are adjusted by R_3 and R_4 so that the grid bias is normal for the 6SN7.

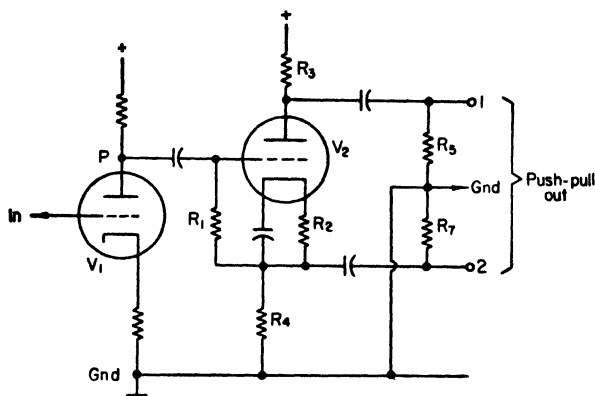



FIG. 53. Split-load (R_3R_4) single-tube phase inverter. No gain from V_2.

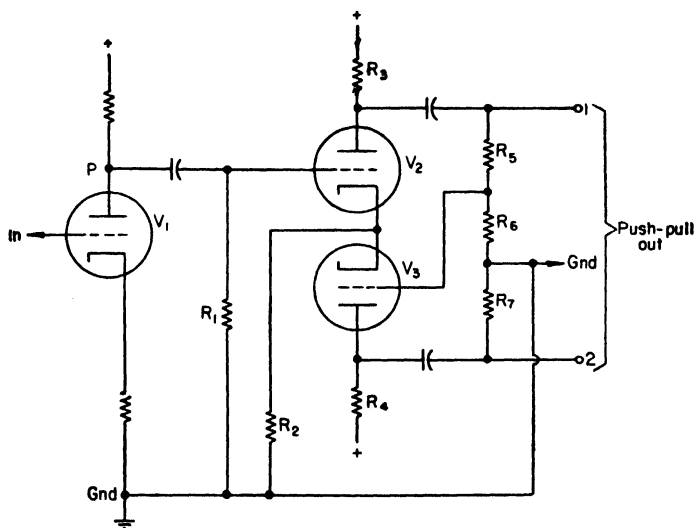



FIG. 54. Tapped-grid resistor (R_5R_6), two-tube phase inverter.

Push-pull outputs are developed at the plate and cathode of this tube because of the matched load resistors R_5 and R_7. To use low-voltage swings, the push-pull 6SN7 tubes provide additional amplification.

Two-tube Phase Splitter. In Fig. 54, V_2 and V_3 are identical and may be a dual triode with a single cathode. The grid resistor for the following stage is R_5 and R_6 in series, with the relations

$$R_5 + R_6 = R_7 \quad \text{and} \quad R_5/R_6 = VG$$

where VG is the voltage gain of V_2 for the voltages and resistors used. Control grid of

V_3 is fed from the junction of R_5 and R_6 with phase-reversed voltage, so that, in turn, the plate of V_3 can feed terminal 2 with reversed voltage (as referred to terminal 1) with respect to ground. V_2 and V_3 each give normal triode gain. R_3 and R_4 should be matched for balanced operation. Omission of by-pass capacitor across the bias resistor R_2 gives better balance in this circuit.

Slightly better high-frequency response can be obtained by separating R_3 into two sections with the same ratio as the voltage gain of the tube, feeding the grid of V_3 via a

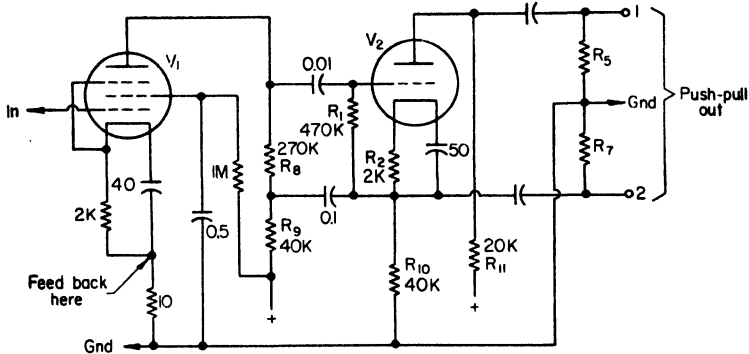



Fig. 55. High-gain phase splitter. R_9 and R_{10} in parallel for audio frequencies match R_{11}.

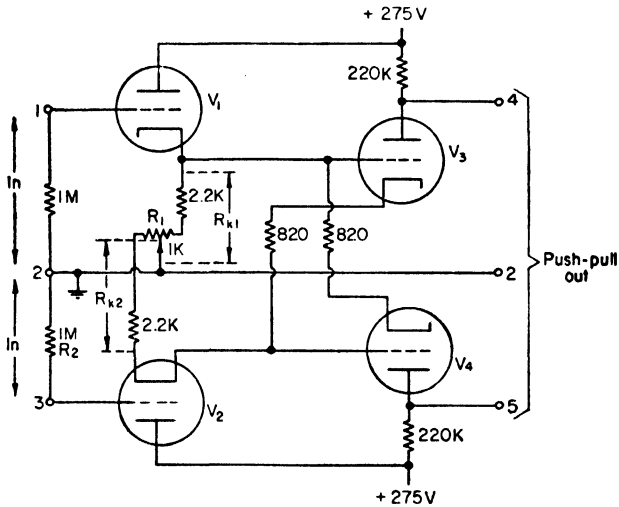



Fig. 56. Cross-coupled phase inverter. R_1 is balance adjuster.

blocking capacitor from the junction, and connecting a resistor to ground from the grid of V_3 for bias.[1]

High-gain Phase Splitter.[2] In Fig. 55 the actual phase splitter V_2 is a triode with less than unity gain, but the pentode driver has six times the usual voltage gain for such a circuit, 1,200 being cited in the reference. Tubes in the original circuit were EF36 pentode driver and triode-connected EF36. Practically the same results may be obtained with type 5879, 6SJ7, or 6J7. The plate resistor R_{11} is matched by R_9 and R_{10}, which are effec-

[1] Cocking, W. T., Push-pull Input Circuits, *Wireless World*, **54**, 85, March, 1948.

[2] Jeffrey, E., Push-pull Phase-splitter, High-gain Circuit, *Wireless World*, **53**, 274, August, 1947; reprinted in "High Quality Amplifiers," Iliffe and Sons, Ltd., London.

tively in parallel for signal frequencies. The maximum output voltage for V_2 is limited because the voltage developed is split into two parts, as in Fig. 53.

Cross-coupled Phase Splitter.[1] In Fig. 56 V_1 and V_2 are low-mu cathode followers with less than unity gain, and V_3 and V_4 are high-mu triodes. Resistor values are shown for 6SN7 and 6SL7 tubes. The control-grid voltage of V_4 is the difference between the output voltage of V_1 and V_2, and the grid voltage of V_3 is equal and opposite to that of V_4. When a small positive voltage is applied between inputs 1 and 2 to the grid of V_1, the slightly reduced signal appears across R_{k1} and no signal appears across R_{k2}. But the grid of V_3 goes slightly positive and the cathode of V_4 goes slightly positive. The latter is equivalent to the grid of V_4 going slightly negative. Thus a balanced signal is developed at the grids of V_3 and V_4, and appears at the output terminals 4 and 5.

Potentiometer R_1 serves to balance the circuit. This is best done with an oscilloscope and an audio signal.

This circuit may be used so as to mix two independent inputs, each input balanced to ground, providing push-pull outputs. If tube and circuit are matched, the same signal applied to points 1 and 3 will cancel, giving no output. One use of this is a high-impedance phonograph connected at 1 and 3; any hum pickup along the line will not produce hum in the output.

If only one input is desired, the other may be used for a tone control,[2] or if one input is connected from 1 to 2, terminals 2 to 3 may be used for negative feedback by changing R_2 to a low value, of the order of 100 ohms, and feeding point 3 from the output transformer secondary.

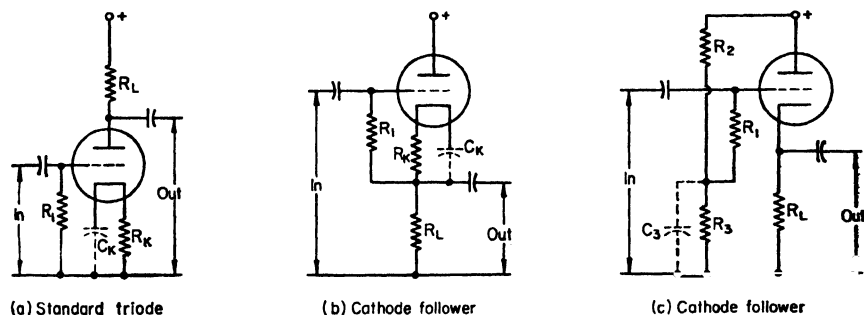



FIG. 57. Standard triode (*a*) compared with cathode-follower stages. In (*c*) bias is obtained from potential divider across plate-voltage supply, thus overcoming some or all of the high negative voltage produced across R_L.

45. Cathode Follower.[3] In a typical cathode-follower stage the voltage gain is 0.8 to 0.95. Its advantages for audio circuits are: (1) its output impedance is low and is largely independent of the circuit which it feeds, (2) it has a very low distortion because of its 100 per cent negative feedback, and (3) it has extremely wide frequency range. Figure 57*a* is a standard triode amplifier circuit with R_L load resistor in series with the plate of the tube. An increase of grid volts decreases the voltage at the plate, and the voltage gain is

$$(E_{\text{out}})/(E_{\text{in}}) = \mu R_L/(r_p + R_L)$$

where μ is the voltage amplification of the tube, R_L the load resistor, and r_p the plate resistance.

Figure 57*b* shows R_L transferred to the cathode circuit. In this circuit the whole tube is shifted toward the positive terminal by the voltage drop across R_L. The

[1] Langford-Smith, F., "Radio Designer's Handbook," 4th ed., Fig. 15.43A, p. 663, RCA, 1953.
[2] VonScoyoc, J. N., A Cross-coupled Input and Phase Inverter Circuit, *Radio and Television News*, **40**, 6, November, 1948.
[3] Albert, A. L., "Fundamental Electronics and Vacuum Tubes," rev. ed., p. 331, Macmillan, 1947. Schultz, T. J., Triode Cathode-followers, *IRE Trans. (Audio)*, **AU-4**, 42, March-April, 1956. Crowhurst, Norman H., Getting the Cathode Follower Straight, *Radio-Electronics*, **25**, 50, December, 1954.

grid bias is supplied by R_k as before. An increase of grid voltage by an input signal brings up the cathode voltage (follower action). The output voltage is less than that of the input because of the 100 per cent feedback. The voltage gain of the cathode follower of Fig. 57*b* is

$$\frac{E_{out}}{E_{in}} = \frac{\mu R_L}{r_p + R_L(1 + \mu)}$$

This is approximately $\mu/(1 + \mu)$ when $\mu > 20$.

The equivalent output impedance is

$$Z_0 = \frac{R_L r_p/(\mu + 1)}{R_L + r_p/(\mu + 1)}$$

when R_L is greater than r_p and μ is greater than 20, approximately

$$Z_0 = r_p/(\mu + 1)$$

In Fig. 57*c*, the junction of R_2R_3 supplies grid voltage above ground and serves to maintain the grid bias point more or less fixed even though the grid takes current on high input voltages. C_3 is optional.

The Miller effect of amplified input capacitance gives a shunting effect for the standard triode circuit of Fig. 57*c*, thus limiting the high-frequency response. The cathode follower has a reduced input capacitance, an inverse Miller effect. This is effective mainly at frequencies of 50,000 cycles and higher, above the usual audio range. A tone-control circuit is a good place to use a cathode follower (see Fig. 14). The cross-coupled phase inverter in Fig. 56 uses two cathode followers; the preamplifier of Fig. 7 uses two cathode followers. Another place for a cathode follower is at the output of a preamplifier, as in Fig. 7.

46. Volume Expander. A volume expander increases the usable dynamic range by providing higher amplification to a large input signal than to a small input signal. This is of interest mainly where soft music is to be played very softly with a minimum of background noise, and loud passages are to be played very loudly. The ideal expander should react quickly to an increase in input-signal level but decrease its amplification more slowly when the signal level goes down. The circuit[1] of Fig. 58 is a single-ended "surgeless volume expander."

This is accomplished by using two cross-connected 6SJ7 tubes, the No. 2 grid and plate of V_1 being connected, respectively, to the plate and No. 2 grid of V_2. Three metal rectifiers are shown: Se_1 and Se_2 are the doubler plate supply; Se_3 provides −48 volts bias.

With no input voltage, the d-c bias across R_2 is applied to the suppressors of V_1 and V_2. With bias set for zero volts the amplification is a maximum and is a minimum at −48 volts. When an audio signal appears across R_1 it is amplified by V_3, rectified and doubled by V_{4a} and V_{4b}, putting d.c. across 470 K filtered by C_1. These are in series with the suppressors of V_1 and V_2, bringing them above the setting of bias as indicated by R_2, increasing the amplification (unless R_2 was already set for maximum amplification). V_{4c} prevents the suppressors from going positive.

When a change in bias is applied to the suppressors of V_1 and V_2, the dynamic change in plate current in V_1 is balanced by an opposite change in screen current in the V_2. Screen-current change in V_1, similarly, is canceled by a plate-current change in V_2. As a consequence, the variation in effective transconductance of V_1 is obtained without current surges in the output.

Attack time is 0.02 sec for 75 per cent of final gain, and decay time is 0.5 to 1.2 sec for fast or slow, using switch *S*. The maximum useful range is 15 db. Maximum signal input is 0.3 volt, maximum output signal 4 volts.

47. Compressor. For the circuit of Fig. 59[2] with a strong input signal the amplification is a minimum; with a weak input signal the amplification is a maximum; the

[1] Butz, A. N., Jr., Surgeless Volume Expander, *Electronics*, September, 1946, p. 140.
[2] Kenney, T. C., KDKA private communication.

output signal is compressed, covering a smaller dynamic range than the input. This keeps the modulation at a desirable maximum, even though the input level drops drastically. This is not to be used where a large dynamic range is desired. For this case switch S is set on 2, for straight-through amplification. Push-pull operation eliminates the surge which is characteristic of most single-ended expanders, though not for Fig. 58.

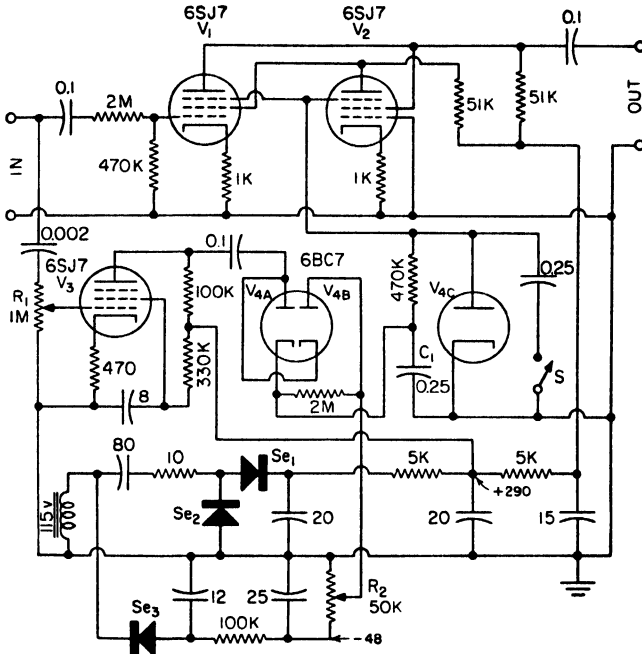



FIG. 58. Surgeless volume expander. Note cross connection of No. 2 grid and plate of the two 6SJ7 tubes, preventing surges in output. V_3 is amplifier with output doubled by two-thirds of V_4 and filtered to provide d.c. across C_1 and, depending on bias from R_2, applied to grid No. 3 of both 6SJ7 tubes. Switch S affects decay time only. Maximum useful range is 15 db. Input 0.3 volt, maximum output 4 volts.

The time delay with components shown is 7 sec; with smaller values of capacitance, the delay will be shorter. This circuit will take care of drops up to 30 db in studio level, with essentially constant output.

When S is in position 1, for low-level input signals, the No. 3 grids of V_1 and V_2 are close to ground, since no voltage is supplied from V_4, T_3, and T_2. Thus maximum amplification is obtained. For high-level input signals, T_2 supplies T_3 with energy which is rectified by V_4 so that the No. 3 grids of V_2 and V_1 are carried negative, by way of the time-delay network. With negative voltage on each No. 3 grid, the 6L7 tubes have minimum amplification. When S is in position 2, each grid No. 3 of each 6L7 is grounded, and the time delay, 6H6, and T_3 are not effective.

The harmonic distortion varies from 1.5 per cent at 50 cycles to 0.4 per cent at 1,000 cycles and 1.7 per cent at 7,500 cycles.

48. Electric Mixing.[1] It is often desirable to control audio volume at some distance away from the amplifier. With "electric mixing" (Fig. 60) there are no variable controls in the signal circuits.

[1] RCA MI-4117 input mixer; MI-4737 compander; MI-4118 master mixer.

From one to four independent inputs may be used, one of which is shown in the figure as Mixer No. 1. Remote volume control P_R is connected to the amplifier circuit by three wires, one of which is ground. Either the internal volume control P_1 or the remote control P_R can vary the d-c bias on No. 1 and No. 3 grids of V_1, a 1612 (a selected 6L7). Thus the controls carry d.c. only and may be run several hundred feet. Three more identical input mixers may be connected at point E.

Between E and F are two capacitors, with a "music" and "speech" selector switch. At F is a dotted connection to an optional compander which may either compress the

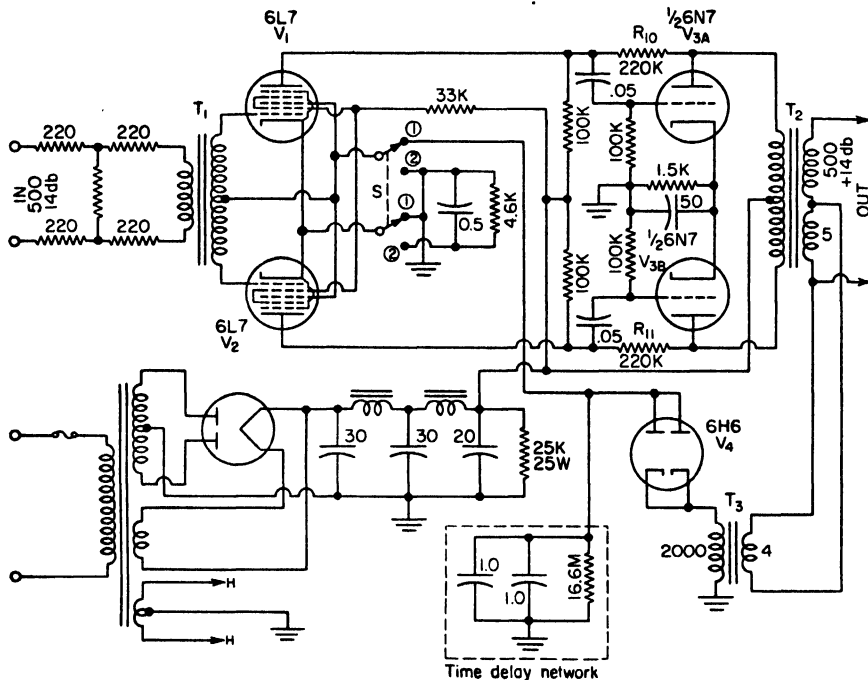



FIG. 59. Maximum modulation amplifier (compressor). With switch S on position 1 changes in studio level will not essentially change the output. Time delay with constants shown in dotted box, 7 sec.

dynamic range or expand it. Expansion is often used with record reproduction to make a larger dynamic range than can be cut into the record. Compression is desirable for speech systems, as it minimizes the varying input, when the talker moves forward and away from the microphone.

Input F is amplified by V_3 doubled by V_4, putting d.c. across the compander volume control; the louder the signal, the more the voltage. When the compander is used the jumper across G and H is removed and the master mixer switch S_2 set in R_2 position.

The master mixer is the final volume control for the input mixers and the compander used. With the jumper in place across G and H, and S_1 at internal position (*i.e.*, not on remote R_2) the compander is not in use and the master is controlled at the amplifier by P_2.

Two-tube Mixer.[1] Five inputs are usable, four at a time, in Fig. 61. There are three independent volume controls: R_1 for magnetic phonograph pickup or high-impedance microphone No. 1 as selected by S_1; R_2 for high-impedance microphone No. 2; R_3 for crystal pickup. The fifth input is a radio which has its own volume control. At maximum gain,

[1] Hitchcock, R. C., Preamplifier-mixer, *Popular Mechanics*, **106**, 141, August, 1946.

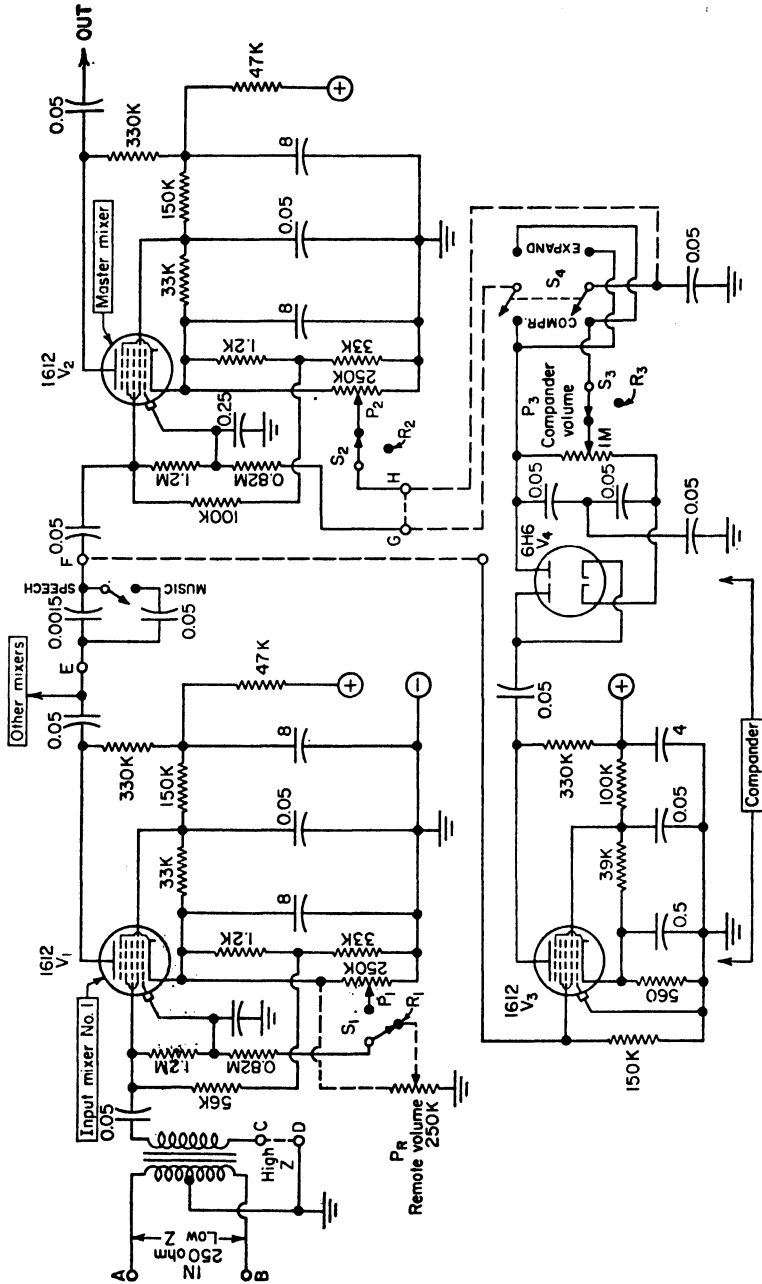



FIG. 60. RCA "electric mixing," shown with one of four possible inputs. With S_1 set on R_1, volume is controlled at a distant point, with d.c., no signal currents being involved. Compander amplifies and doubles the audio signal. S_3 can be set to expand or compress, as desired.

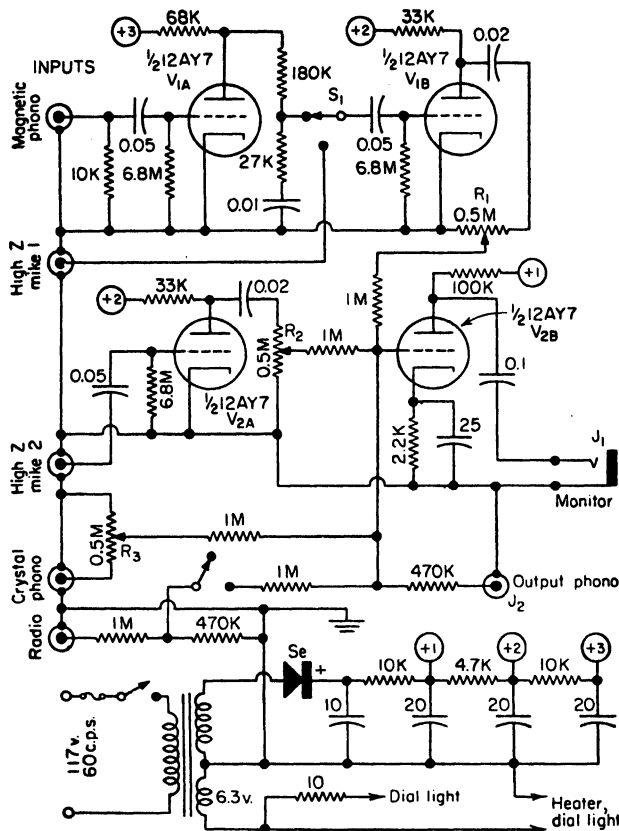



FIG. 61. Mixer for five inputs.

the inputs required to supply 0.05 volt at a 0.5-megohm load at J_2 or 2.0 volts across 0.5 megohm at J_1 are:

	Volts
Magnetic phonograph pickup	0.003
Microphone No. 1	0.01
Microphone No. 2	0.01
Crystal phonograph pickup	0.5
Radio	1.0

These are relative; thus with 0.01 volt instead of 0.05 volt output, each input would be one-fifth the values given.

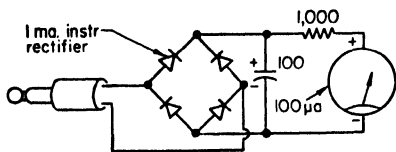



FIG. 62. Monitor voltmeter. Full scale = 0.4 volt. Approximately 10,000-ohms per volt.

V_{2b} is the monitor amplifier which feeds J_1 where a 0.4-volt a-c rectifier meter (10,000 ohms per volt); Fig. 62 is the output level meter. Jack J_1 may also be used with monitor headphones without affecting the gain of the circuits feeding jack J_2. J_2 supplies the combined input signals to the microphone input of a recorder or amplifier. The regulation at J_1 is low; the voltage output with 0.5-megohm load is 2.0 volts maximum but with the voltmeter load it is 0.4 volt.

Originally designed as an input mixer for tape recorder using J_2 as output and a-c monitor

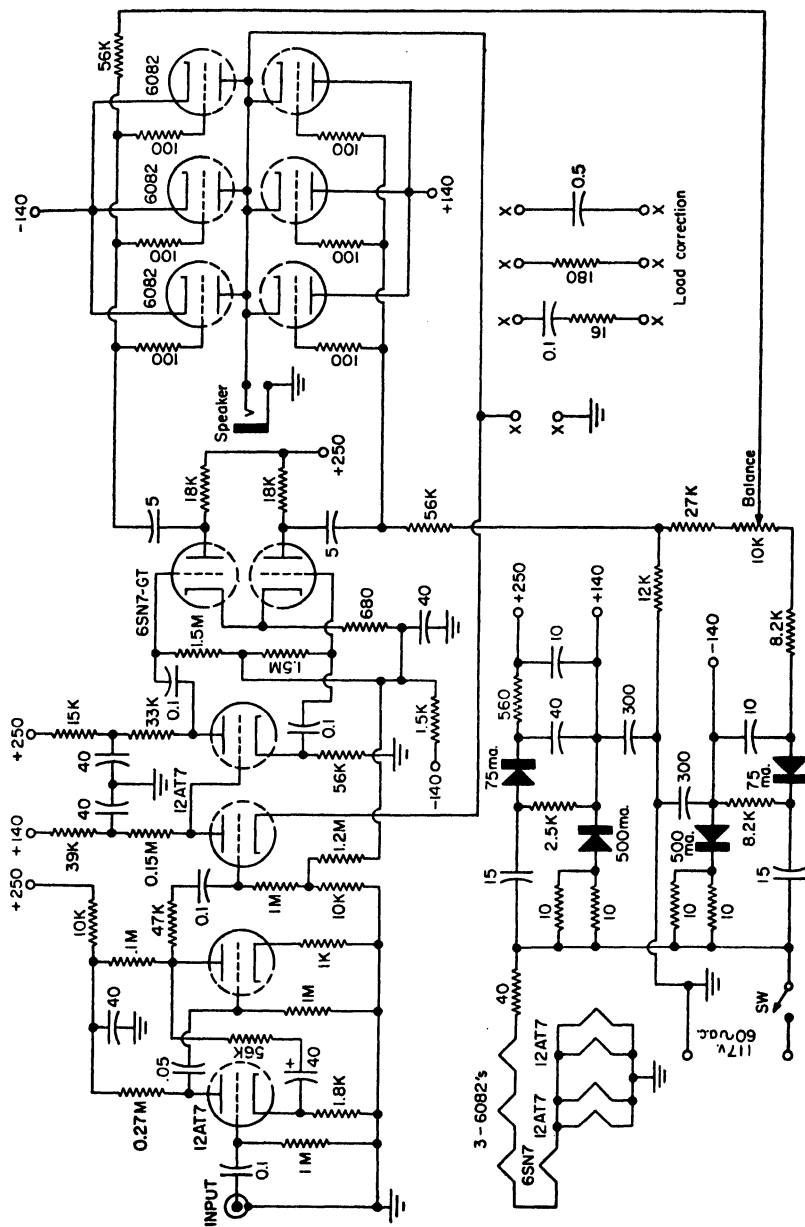



FIG. 63. Transformerless amplifier.

voltmeter at J_1, it can also be used as a preamplifier for a low-gain power amplifier, such as a crystal phonograph amplifier, which requires 0.5 to 2.0 volts input. In this case J_1 output is used.

The self-contained power supply is a 50-ma half-wave selenium rectifier and *RC* filter.

49. Circuits without Output Transformers.[1] The most critical component in an audio amplifier is the output transformer, and various circuits have been suggested to eliminate it. The major obstacle is the high impedance of the output tubes as compared with the low impedance of the loud-speakers. In general, if the output transformer can be eliminated, both ends of the audio spectrum are improved, because of the elimination of the necessity of high primary inductance for low frequencies and low leakage reactance between windings for high frequencies. Circuits without output transformers are sometimes termed "OTL," or "output-transformer-less."

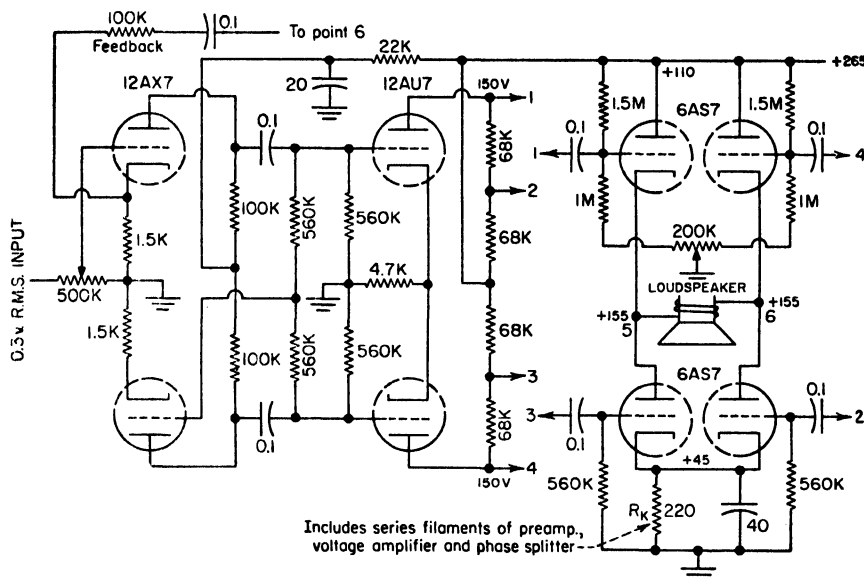



FIG. 64. Transformerless amplifier. Note bridge-type circuit to match voice-coil output: 8 watts into 400-ohm voice coil.

1. Three 6082 dual-triode power tubes comprise the output stage in Fig. 63.[2] A double-ended half-wave rectifier furnishes plate power 140 volts above and below ground. The driver is a 6SN7GT with 250-volt doubler supply. With this circuit the authors obtain 25 audio watts into a 16-ohm speaker with 0.3 per cent harmonic distortion. Negative feedback of 40 db is applied from the speaker voice coil to the cathode of the phase-splitting driver 12AT7. Three possible shunt connections *XX* are suggested to stabilize speaker coils having a high inductance. Input required is 1.0 volt.

2. By using two 6AS7G output dual triodes, a type of bridge circuit[3] drives a 400-ohm speaker with 8 watts at 7 per cent intermodulation distortion without feedback, and 0.8

[1] Peterson, A. P. G., A New Push-pull Amplifier Circuit, *General Radio Experimenter*, **26**, October, 1951. Peterson, Arnold, and D. B. Sinclair, A Single-ended Push-pull Audio Amplifier, *Proc. IRE*, January, 1952; General Radio Reprint A-56. Yeh, Chai, Analysis of a Single-ended Push-pull Audio Amplifier, *Trans. IRE (Audio)*, **AU-1**, 9, March–April, 1953. Amemiya, Hiroshi, Analysis of Drivers for Single-ended Push-pull Stage, *Trans. IRE (Audio)*, **AU-3**, 162, September–October, 1955; correction on p. 27, **AU-4**, January–February, 1956.

[2] Dickie, D. P., Jr., and A. Macovski, A Transformerless 25-watt Amplifier for Conventional Loud-speakers, *Audio*, **38**, 22, June, 1954.

[3] Onder, Kerim, Audio Amplifier Matches Voice-coil Impedance, *Electronics*, **27**, 176, February, 1954.

per cent with feedback. The output stage has four driven points, and a special plate circuit driver 12AU7 driver is employed (Fig. 64). A similar circuit[1] uses four 6082 dual triodes, and 18 watts is obtained with a load of 200 ohms.

3. Several circuits[2] use parallel 12B4 miniature noval power triodes for the output stage, with 157 volts d.c. above and below ground for plate power (Fig. 65). With eight 12B4 tubes an output of 10 watts into a 16-ohm speaker is obtained, and with 14 such tubes, 20 watts into 16 ohms. The driver is a 6S4 split-load inverter with 400-volt separate d-c supply.

50. Synthetic Bass. By intentionally supplying only the high harmonics of a bass tone, the ear will reconstruct the tone, even though the fundamental is not

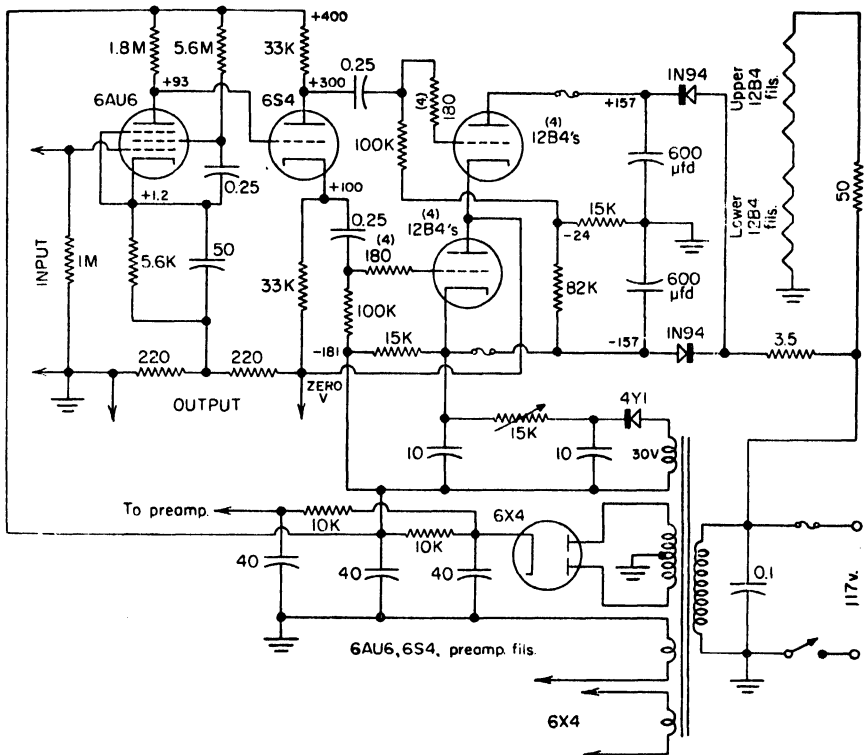



FIG. 65. Audio amplifier with no output transformer. Note output at lower left. Using eight 12B4 tubes in two groups of four, 10 watts is provided for a 16-ohm speaker.

present.[3] For table radios and other devices using small loud-speakers where no provision for baffles or boxes is possible, this scheme permits a "synthetic bass" which extends the low range an octave or two below 300 cycles although a critical musical ear will detect the excess of higher harmonics in the synthetic low.

In Fig. 66 C_1 has a relatively small value and in conjunction with R_1 will progressively reduce the passage of frequencies below 200 cps. The R_3C_2 circuit supplies strong positive feedback from V_2 to V_1 at 80 cps with decreasing feedback above this value. An incoming

[1] Onder, Kerim, A New Transformerless Amplifier Circuit, *J. Audio, Eng. Soc.*, October, 1953, p. 282.

[2] Futterman, Julius, An Output-transformerless Power Amplifier, *J. Audio, Eng. Soc.*, **2**, 252, October, 1954.

[3] Hitchcock, R. C., Demonstrating Harmonics and Beats, *Am. J. Phys.*, **19**, 445, November, 1951.

wave of 4,000 cycles will be amplified in V_1 and passed on to V_2 with practically no attenuation, but practically none of it is fed back to V_1. However, V_2 amplifies this 4,000-cycle wave and passes it on to the speaker.

Consider now a frequency of 80 cycles. This is amplified in V_1, attenuated by C_1, R_1 and amplified by V_2. It is selectively regenerated by C_2 and R_3 and again amplified by V_1 and V_2. The curve of the output vs. input is steeper for low intensities, and this curve is accentuated when V_1 is a remote-cutoff pentode, thus encouraging odd harmonics, notably the third and fifth, which cause the ear to reconstruct the bass tone.[1]

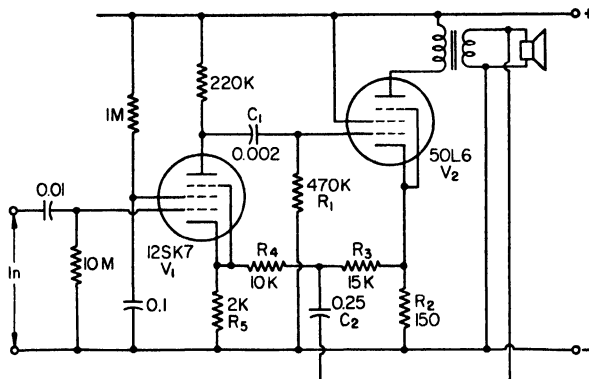



FIG. 66. Synthetic-bass amplifier. Intentional distortion is produced by using curved characteristics of tube. Note high-frequency C_1R_1 network feeding output tube, and strong positive feedback for low frequencies via R_3R_4.

TEST EQUIPMENT

51. Voltmeter-Ammeter. A multirange d-c voltmeter is useful for checking grid bias and plate and screen voltages, both load and no-load. A cardinal principle of measurement is that the measured quantity should be changed as little as possible by the measuring device. If the current through a resistor is 1.0 ma, the voltmeter which is connected in parallel with it should take a current much smaller than 1.0 ma. As a rule, a 200-μa (or less) full-scale voltmeter is desirable, with 5,000 ohms per volt resistance or more.

Circuit diagrams often show the nominal voltages to be expected, and checking these values within 20 per cent will offer a quick analysis of the circuit. Tube testers are seldom as desirable as voltage checks of the tubes in actual service.

Figure 67 shows a two-tube amplifier with d-c voltmeters and d-c milliammeters at possible test points. Seldom will all of these be required or used.

The general plan for testing a new circuit is: find out the voltage supplied E_7. Check the power pentode V_2 first; find its grid No. 1 bias E_6. Since plate currents are usually more than ten times grid No. 2 currents, I_6 is approximately equal to I_5, and hence I_5 is very nearly the same as E_6/R_6 amp. E_{44} No. 2 grid volts should be checked; then if No. 1 grid bias is about right for the supply voltage E_7, and I_5 normal, V_2 and its circuit can be assumed satisfactory to use.

To connect milliammeters requires breaking a circuit and bringing two leads to the meter; this is not attractive because both leads have to be insulated for the voltages encountered. Insulated test clips from a voltmeter are readily attached, and usually tell enough of what is wanted to give a true operating picture of the d-c circuit. For example, $I_2 = E_2/R_2$.

52. Ohmmeter. A battery-powered ohmmeter is a useful test tool. It can be used to check circuit components: (1) resistors and potentiometers; (2) primaries

[1] Shepard, F. H., Jr., U.S. Patent 2313098, Fig. 2, Mar. 9, 1943.

and secondaries of transformers; (3) leakage of capacitors, electrolytic or paper; (4) grounds on components; and (5) short circuits.

Ohmmeter ranges are often given as "full scale" such as 1,000/10,000/100,000/1,000,000 ohms. Ohmmeter pointers normally indicate zero ohms with the pointer fully deflected toward the right, and are set at that point by an adjusting knob when the two test leads are touched together. The highest ohmages, so-called "full scale," are toward the left, with the pointer moved just slightly from the left-hand no-reading position. The precision of an ohmmeter reading is neither high nor constant, being most accurate up scale and least accurate down scale. However, it is often sufficient to determine the suitability of a component for use.

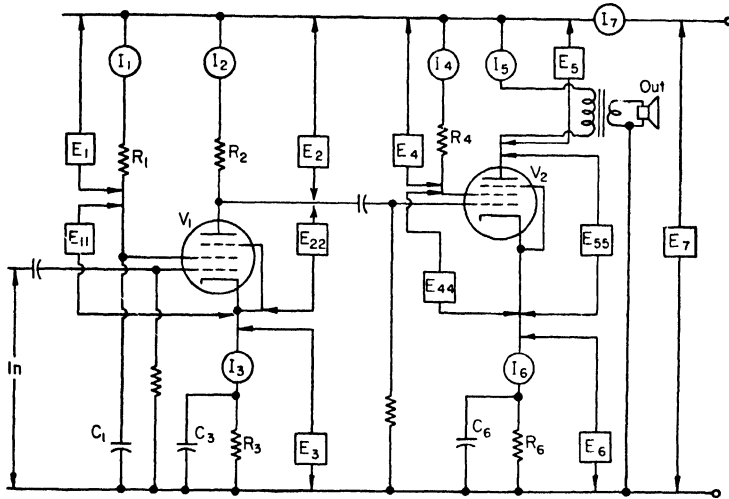



FIG. 67. Use of d-c meters to check amplifier.

To check for resistance, the component ordinarily cannot be in an energized circuit. The power supply must be cut off, and usually all but one of the terminals to the component disconnected.

Notes. Potentiometers with three leads should have at least two leads disconnected, to read ohms without including other components. Resistors with two leads can be read for ohms with only one lead disconnected. Multiple-lead resistors must have all but one lead disconnected.

The d-c resistance of transformer windings may be found with an ohmmeter, but there will be an inductive kick as the circuit to the ohmmeter is broken.

An electrolytic capacitor has appreciable resistance in either direction, but it is much lower in its reverse direction. When positive voltage from the meter is applied to the plus terminal of the electrolytic, this resistance will be the higher of the two. Some volt-ohm-milliammeters have a reversal of polarities; the positive lead for volts and milliamperes is the negative lead when used as an ohmmeter. Paper capacitors should show practically no measurable steady d.c. in either direction. There will be an initial charging kick which quickly dies to zero. A paper capacitor showing a steady deflection of 1 megohm or less is defective.

Checking resistance of various parts of the ground is often informative.

A short-circuited component reads zero on an ohmmeter. When checking resistances greater than 10,000 ohms, do not touch the ohmmeter leads. Dry-skin resistance of 50,000 to 100,000 ohms is typical, and wet hands may give a reading as low as 10,000 ohms. In placing the ohmmeter terminals, hold them by the insulated handles, on a deenergized circuit.

53. Vacuum-tube Voltmeter. These instruments usually have several ranges and are very helpful for measuring a-c and audio voltages. To measure interstage audio voltage, a series capacitor of 0.002 to 0.02 μf will prevent changing the d-c conditions. Most VTVMs read peak voltages but are calibrated in rms, assuming sine waves. This makes them read wrong for waves other than sine (see Fig. 68). Their input resistance is usually high, from 0.5 to 10 megohms, and their connection to low-energy circuits does not materially affect circuit operation.

54. Cathode-ray Oscilloscope. This is one of the most versatile of test instruments. It has high input resistance so that it does not affect the circuits across

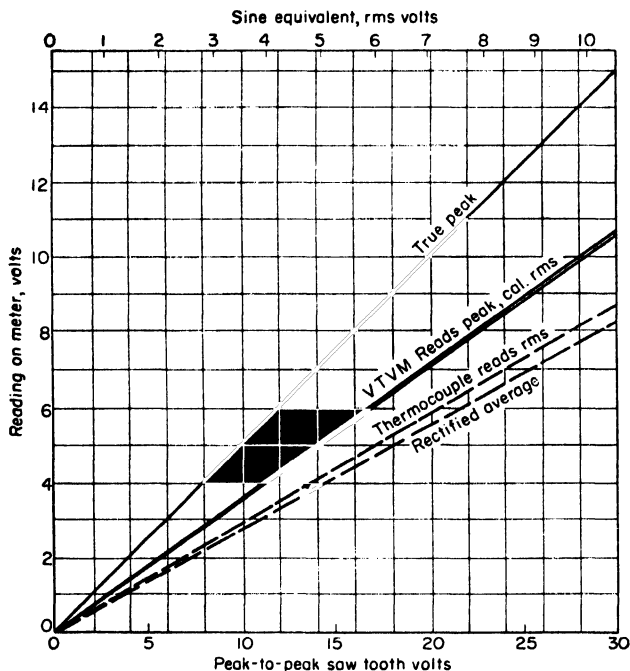



FIG. 68. Meter indications for saw-tooth waves.

which it is connected. Linear time scanning of the beam from left to right and vertical deflection by the voltage of the circuit being studied are most often used. Time scanning from 10 to 10,000 per second includes practically all audio-amplifier work.

Connecting a cathode-ray oscilloscope across the amplifier output and watching for peaks of the amplified waves will show immediately if there is any clipping. If the clipping is in the push-pull output stage alone, the peaks will be clipped at top and bottom. If a single-ended stage is clipping, usually either the top or the bottom is flattened.

A saw-tooth oscillator (Sec. 57) plus a cathode-ray oscilloscope can give a variety of important test data (see Figs. 72 and 73): (1) flat frequency response, sharp corners, and straight lines; (2) treble boost, an extra high starting spike; (3) treble cut, rounded starting peak; (4) bass boost, a bulged center; (5) bass cut, a concave center; and (6) overload due to clipping, flat tops in saw-tooth wave.

A cathode-ray oscilloscope clearly shows gross distortion of sine waves, but considerable practice is needed to detect a small amount of distortion in a sine wave.

55. Equivalent Single-sine-wave Power. Audio amplifiers are rated at their maximum electrical output when delivering a single sine wave. Power amplifiers are rated two ways: the wattage rating for continuous sine-wave power output, and the "audio" rating. The audio rating is about twice the continuous rating. Considered here is the sine-wave rating.

To find the frequency response, sine waves of different frequencies are supplied one at a time, and the output is measured across the rated load resistor. The power input to the amplifier $P_{in} = (E_{in})^2/R_{in}$; and the power out $P_{out} = (E_{out})^2/R_{out}$, where the Rs are in ohms, the Es in volts, and the Ps in watts. The over-all response, at any specified frequency, is

$$\text{db} = 10 \log (P_{out}/P_{in})$$

Now when two sine waves are simultaneously applied, as in intermodulation testing (Sec. 56), the actual P_{out} requires a conversion factor to relate it to the equivalent single-sine-wave power. Suppose two output voltages are E_1 and E_2 and that their ratio is n so that $E_2 = nE_1$, and both are impressed on the same load resistor; the total power in the load R is

$$P_{1,2} = (n^2 + 1)E_1^2/R$$

In testing the intermodulation, $P_{1,2}$ is the actual power received by R. Note, however, that, when two waves are applied simultaneously, their peaks add together. Clipping these peaks is one cause of intermodulation distortion. In the example mentioned, the new peak corresponds to a new rms voltage $(n + 1)E_1$. The power in a resistor with this voltage would be

$$P_1 = (n + 1)^2E_1^2/R$$

So when an amplifier is tested with two waves E_1 and nE_1, giving $P_{1,2}$ power with no clipping, it really can handle P_1 single-sine-wave power, with the relation

$$P_1 = P_{1,2}(n + 1)^2/(n^2)$$

For intermodulation testing, where $n = 4$,

$$P_1 = P_{1,2}(25/17) = 1.47P_{1,2}$$

That is, the equivalent single-sine-wave power is 1.47 times the two-sine-wave power actually in the load resistor R, for a 4:1 voltage ratio or 12 db.[1]

The most accurate a-f voltmeter under all conditions is the thermocouple type because it reads true rms even for waves of different shapes. However, thermocouples are seldom used, except for checking other meters, because of their low overload capacity. Rectifier-type meters actually read rectified average, although they are calibrated to show rms sine-wave voltages. They are reasonably accurate for sine waves but the form factor of some peaked waves requires recalibration. VTVMs are peak-reading meters but also may be calibrated for sine waves. Waves other than sine will not read correctly without conversion factors. See Fig. 68 for sawtooth and the readings of these meters.

56. Measuring Intermodulation Distortion.[1] In Fig. 69 the power amplifier being tested has an output load resistor R. Meter E_{out} shows the output voltage. The input generator produces sine waves, each of controlled amplitude E_{100} and $E_{5,000}$, whose combined value is E_{in} volts. An optional but informative instrument is the oscilloscope CRO driven by an auxiliary amplifier if necessary. A quadruple bus is connected to the two four-point switches S_1 and S_2 allowing the use of the auxiliary amplifier or the VTVM direct as required.

The filter indicated on Fig. 69 is made up of a high-pass section passing 5,000 cps and a

[1] Aston, R. H., Intermodulation Measurements on Radiotron Amplifier A515, *Radiotronics*, No. 130, p. 34, March–April, 1948. Hitchcock, R. C., Intermodulation Distortion, *Audio Eng.*, **36**, 21, October, 1952. Various combinations of frequencies for IMD measurements have been suggested. For example, (1) 60:7,000; 40:12,000 cps, Standard Transformer Corp.; (2) 40:2,000; 40:7,000; 40:12,000 cps, Acro Products Co.; (3) 40:4,000 cps, Thomas Roddam, *Wireless World*, **56**, 122, 1950; (4) 60:4,000 cps, Pierce J. Aubrey, *Audio Eng.*, **35**, 22, 1951; (5) 100:7,000 cps, Cook Series 10 Record.

pi-type low-pass section passing 100 cps. Both sections are down 20 db at 1,000 cps, the two sections being isolated by a 1-K resistor across which is shun'ed a diode. A shunt resistor terminates the filter across which the distortion E_{LO} appears and is read.

An input signal E_{in} of, say, 0.4 volt E_{100} and 0.1 volt $E_{5,000}$ is applied to the amplifier, and the output E_{out} is read. E_{out} enters the filter giving E_{HI} and emerging as E_{LO}, which is read by the VTVM for one value of the input. For an ideal amplifier, E_{LO} is zero. For a real amplifier

$$\% \text{ intermodulation distortion} = 100E_{LO}/E_{HI}$$

for any given E_{out}, that is, for a given output power. The equivalent single-sine-wave power in the resistor is $1.47(E_{out}^2)/R$.

Figure 70 is a set of typical graphs[1] of E_{in} volts against E_{out} (measured), W watts (calculated), E_{HI} (measured), IMD intermodulation distortion per cent (calculated), and E_{LO} (measured) for a Williamson amplifier, using a Savage 2B36B output transformer, and an 8-ohm resistance load R.

Fig. 69. Circuit for intermodulation distortion test. Dotted circles are optional meters. Two-tone sine generator provides input. Voltage output is connected to filter. With no intermodulation distortion E_{LO} will be zero.

The distortion E_{LO} rises slowly at first, though at a faster rate than E_{HI}, showing that the intermodulation distortion is rising. Then E_{LO} rises at a still faster rate, finally tapering off.

Intermodulation-distortion figures above 5 per cent indicate a poor degree of fidelity. This amplifier, rated at 15 watts, had 0.4 per cent intermodulation distortion at that point, and with 20 watts output it had 1.3 per cent. Note that the VTVM must read rather low voltages. E_{LO} is only 0.0001 volt (0.1 mv) at the lowest reading on the figure.

57. Saw-tooth Generator. A saw-tooth generator (Fig. 71) includes all harmonics.[2] Each harmonic is smaller in amplitude than the one preceding it, the nth harmonic being $1/n$ that of the fundamental. Saw-tooth waves are more desirable for testing than square waves, since square waves contain only odd harmonics. Square waves cannot be used to show clipping of peaks, since they are already flat-topped.

The steep wave front of a saw-tooth will start an amplifier ringing if there is any tendency to do so; see Sec. 62 and Fig. 77. As corrective measures are taken, the oscilloscope will show the improvement, and when the final saw-tooth is sharp, the correction is complete (see Fig. 77).

The saw-tooth generator can also show overload due to clipping. The amplifier

[1] Hitchcock, R. C., Oscilloscope Pictures of Intermodulation Distortion, *Am. J. Phys.*, **22**, 187, April, 1954, Figs. 1 and 6 (recalculated).

[2] Hitchcock, R. C., Testing with White Sound, *Audio*, **38**, 41, October, 1954. For methods of measuring gain and loss and frequency response of audio amplifiers, see *Proc. IRE*, **44**, 668, May, 1956, IRE Standards on Audio Systems and Components.

is driven with an increasing saw-tooth voltage while the oscilloscope displays the output waves. When the tops just begin to flatten, the maximum peak-to-peak voltage-handling ability of the amplifier has been reached. From this the equivalent single sine wave can be determined. The conversion requires a knowledge of the meter being used. Thus in Fig. 68 note that a peak-to-peak saw-tooth voltage of 20 has the same maximum and minimum as a sine wave having an rms value of 7.07 volts. At 20 volts peak-to-peak bottom scale and 7.07 volts rms top scale the rectifier meter shows 5.6 volts, the thermocouple 5.8 volts, and the VTVM 7.07 volts. If this peak-to-peak saw-tooth of 20 volts is the greatest value which the amplifier will produce without clipping, the equivalent-sine-wave power is E^2/R, where E is the equivalent-sine-wave volts; the power then is $(7.07)^2/R$, where R is the terminating resistor being used.

FIG. 70. Typical intermodulation distortion (IMD) of Williamson amplifier. Solid lines show measured quantities, E_{LO} from filter output, E_{HI} from high-pass filter, E_{out} amplifier output volts on 8-ohm load. Dotted lines are calculated: output power and distortion.

Using a resistor for a load is standard practice, and quiet. It is the first step in

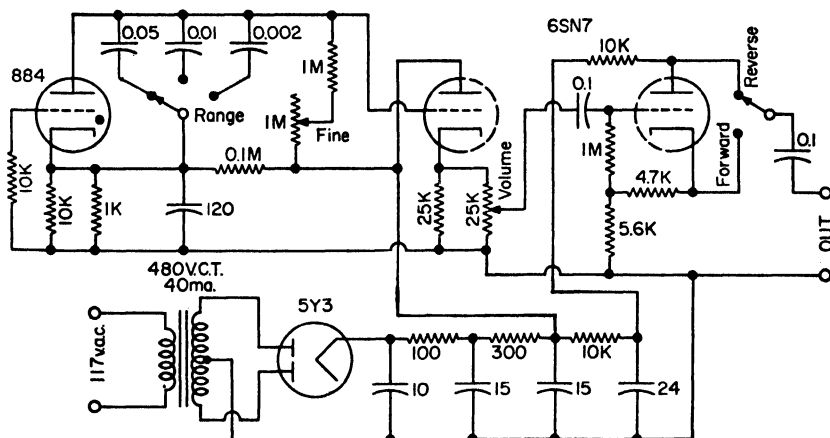



FIG. 71. Saw-tooth-wave generator. Note reverse and forward switch. Nominal ranges are 100/1,000/10,000 cps. Output is 20 volts peak to peak into loads of 100,000 ohms or higher.

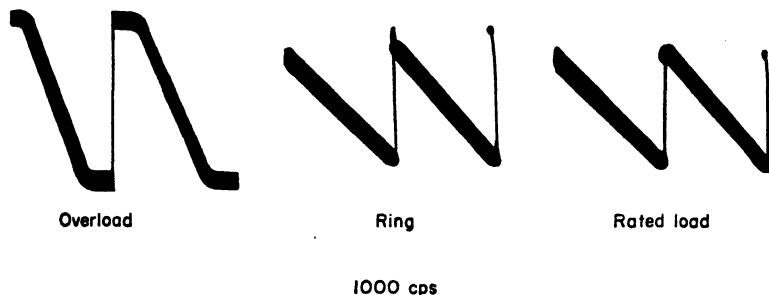



FIG. 72. Use of 1,000-cps saw-tooth in amplifier test, indicating overload (clipped tops and bottoms), ring (spike at top), rated load (sharp saw-tooth).

testing a new amplifier. But before the amplifier can be considered fully checked, the actual output circuit should be connected. Sometimes crossover networks provide unexpected results. The test, however, can be made at lower power, and the output voltage watched on the oscilloscope as the saw-tooth input voltage is increased. A quick way to find out the flat setting of the bass and treble controls is to have the saw-tooth come out sharp, with straight lines. A 1,000 per second saw-tooth will check bass down to 100 cps and treble up to 10,000 cps.

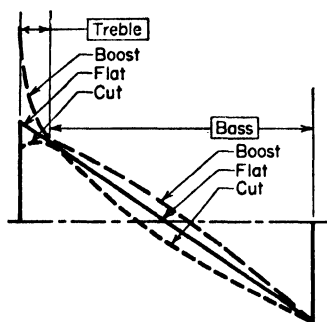



FIG. 73. Saw-tooth-wave interpretation. Treble response, 10 per cent; "boost" shows a spike, "flat" a sharp corner, "cut" as a rounded corner; bass, 90 per cent.

Saw-tooth Oscillator Circuit. The saw-tooth oscillator in Fig. 71 uses an 884 gas triode. The nominal ranges are 100/1,000/10,000 cps with the fine adjustment being a 1-megohm variable resistor in series with a 1-megohm grid resistor of a 6SN7 cathode follower. The final 6SN7 is a forward-reverse tube, having a split load. For single-ended amplifiers the saw-tooth should be tried both forward and reverse, as sometimes these amplifiers act differently when the saw-tooth goes negative first, instead of positive first as shown in Fig. 73. Into a load of 0.1 megohm or higher, this generator supplies up to 20 volts peak to peak.

58. Measuring Harmonic Distortion. When a single sine wave is fed to an amplifier input, the output contains the fundamental (first harmonic) plus a smaller amount of second, third, and higher harmonics. These unwanted harmonics are present to

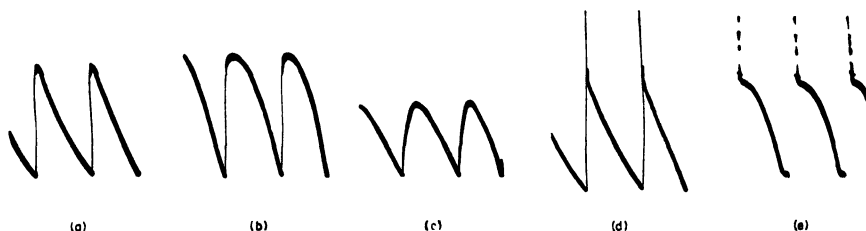



FIG. 74. A 1,000-cycle saw-tooth shows amplifier performance when boost and cut are at 100 and 5,000 cps. (*a*) At 100 cps cut is 6 db. (*b*) At 100 cps boost is 14 db. (*c*) At 5,000 cps cut is 12 db. (*d*) At 5,000 cps boost is 8 db. (*e*) Bass boost (100) is 14 db, and treble boost (5,000) is 8 db.

some extent at all operating levels, but in general they will be greatest at the full output of the amplifier.

There are two methods of testing for harmonic distortion. Both apply a low-distortion sine wave to the input. Next the single frequency is balanced out of the output, by using a tuned bridge circuit. At this point the simpler of the two methods measures the residual output, the "total" harmonic distortion.[1] The fractional distortion in per cent is the residual divided by the full output voltage, multiplied by 100.

A more accurate method, after balancing out the fundamental at the output, is to tune in the second, third, and higher harmonics, one at a time by an additional bridge, and to measure each separately. A commercial wave analyzer for this purpose is shown schematically in Fig. 75. This is a high-gain amplifier with *RC* interstage coupling, balanced to a sharp null, a calibrated attenuator for adjusting sensitivity, and a built-in VTVM. Distortion as low as 0.1 per cent can be measured for third harmonics of fundamentals as high as 15,000 cycles and higher harmonics of lower frequencies. In addition a low-distortion sine-wave generator is needed.

Since it is usual for most audio amplifiers to

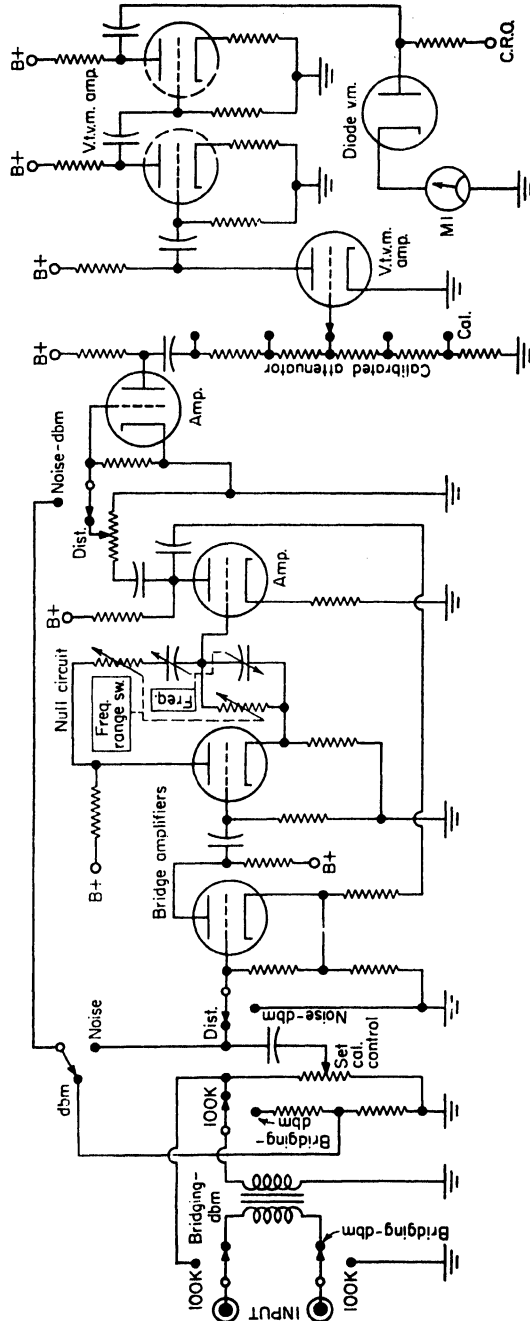



FIG. 75. Wave-form measuring circuit, used to measure harmonic distortion, made up of a high-gain amplifier, null circuit, calibrated attenuator, and VTVM. A pure sine-wave generator input is needed. (*Courtesy of General Radio Company.*)

[1] Tyler, V. J., Simple Distortion Meter, *Wireless World*, **49**, 431, September, 1953.

have a total harmonic distortion of less than 10 per cent, an approximate working formula is[1]

$$\% \text{ total distortion} = \sqrt{(k_2)^2 + (k_3)^2 + \cdots (k_n)^2}$$

where k is the per cent of the voltage at the harmonic specified by the subscript, to the output voltage.

AMPLIFIER TROUBLES

59. Hum. The possibility of 60- and 120-cycle power-line hum is always present in an audio amplifier. Perhaps it is most serious for home high-fidelity sets, where quiet surroundings enhance any extraneous sounds.[2] Commercial amplifiers usually specify the amount of hum, for example, for the amplifier of Fig. 25 as "80 db below full output," which in this case is 30 watts. Putting h as the hum decibels, P_{out} the rated power, and P_{hum} the hum power,

$$h = 10 \log (P_{out}/P_{hum})$$
$$P_{hum} = (10^{-h/10}) P_{out} = 10^{-8} \times 30 = 0.3 \times 10^{-6} \text{ watt}$$

The problem of hum can be divided into two categories: the design and construction of the amplifier, and (2) its operation and use.

1. In designing a high-quality amplifier, a schematic circuit is not enough. The actual point-to-point wiring should be shown on a scale drawing. Grid and plate circuits should be separated from each other and from a-c heater wires; high-gain input tubes should be as far as possible from the power tubes, and especially from the rectifier tube. For a "ground" a large tinned copper bus bar is advisable, connected to the metal chassis at one point. Usually the best place is at the input jack positions. No closed loops in the ground bus should be permitted. Loops may include magnetic flux from the power transformer and thus pick up hum. Instead, the bus should run from one end of the chassis to the other, with radiating jumpers to provide grounds as needed. Grounding to chassis at many points is bad practice.

Metal tubes are preferable to glass tubes for input stages, though shielded miniature tubes are satisfactory. For an input amplifier, a 6.3-volt heater supplied with 5.8 volts will give less hum. For dual voltages, 12.6 or 6.3 volts, the parallel 6.3-volt connection will produce minimum hum on a.c. With d.c. on the heater, the 12.6-volt connection is satisfactory (see Fig. 25).

2. Properly connecting the outputs and inputs to an amplifier can reduce hum. The way to start is to connect the output circuits (no inputs). Turning up the amplifier controls, bass boost, and volume should give very little or no noticeable hum. The only devices in the output circuit susceptible to hum are the air-core crossover coils. These coils should be placed several feet away from the power transformer, and not too close to the phonograph motor; both are potential sources of hum. The polarity of the power-line plug will affect the hum of an a.c.-d.c. amplifier such as Fig. 22, and to some extent any high-gain amplifier. The plug should be turned over and reinserted to find out the minimum hum.

The metal chassis of a high-gain amplifier, fed from a microphone or a magnetic pickup, will often have considerably less hum when grounded. Water-pipe grounds are best, with hot-water-heating-system pipes next best. Some tubes have more hum than others. Exchanging similar-type input tubes, to find the arrangement with least hum, is worth trying.

If more than one input is to be used, they should be connected one at a time, noting the hum for each as the volume is turned up. Leads to all inputs should be shielded cables grounded to the amplifier chassis. Input leads, even if shielded, should be kept away from the two main sources of hum, the power transformer and the phonograph motor.

60. Distortion. Gross distortion in an audio system is due to either external or internal causes. Mismatching at input or output is a frequent external cause. A crystal phonograph pickup with 0.5 volt output, if connected to a magnetic phono-

[1] Conway, T. D., Measurement of Harmonic Distortion, *Wireless World*, **60**, 110, March, 1954.

[2] Fleming, L., Controlling Hum in Audio Amplifiers, *Radio and Television News*, **44**, 55, November, 1950.

graph input, will overload the input stages of the amplifier. Since the volume control is farther along in the circuit, the voltage cannot be reduced at the input. The remedy is to use the proper input jack, or to reduce the 0.5 volt by a two-resistor voltage divider to about 0.01 volt.

Mismatching outputs has a similar effect. An 8-ohm speaker on a 200-ohm tap will not produce proper and expected volume, and the frequency range will be restricted. In driving the amplifier harder to get more output, the final stage will undoubtedly be overloaded, with bad distortion. Usually for public-address work the match from 75 to 125 per cent of the rated impedance will be close enough.

A simple trouble has a simple answer: the output terminal strip often has connections so close together that a wire may touch two adjacent terminals. This short circuit greatly reduces an amplifier output, and in trying to get sufficient volume, the amplifier is overloaded and the speaker shows it.

Internal distortion can occur because of lack of sufficient control-grid bias on the power-output tubes, which allows the plate current to increase and may cause rectifier-tube plates or the output-tube plates to get red hot.

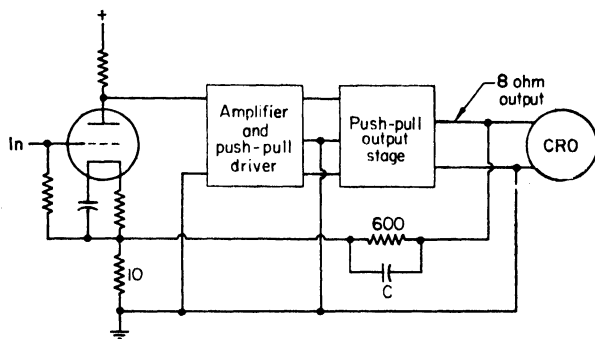



Fig. 76. Amplifier which rings, as shown in Fig. 77, has capacitor C added across the negative-voltage feedback 600-ohm resistor.

Harmonic and intermodulation distortion are more complex and are not in the class of gross distortion.

61. Noise. Noise is any unwanted sound in an amplifier. Power-line hum is one kind. With no input connected and amplifier gain at maximum, a frying or hissing sound is usually due to a tube, resistor, or capacitor. With a microphone input, the background noise and microphone hiss usually mask any internal noises. Tubes may have leakage from heater to cathode, a resistor in a preamplifier may run too hot, or a paper capacitor may leak.

Microphonics are a special type of noise. Usually an input tube will "ring" when tapped with the finger. If the sound dies away quickly, for some installations this is not serious, but since tubes vary in this property, a better tube can probably be found and used. A high-power loud-speaker should not face or be placed too close to a high-gain amplifier input tube, nor should the speaker be too close to a low-output phonograph pickup; pickups are microphonic.

62. Oscillation and Feedback. Oscillations are unwanted frequencies in the amplifier; they may be above audible, audible, or below audible in frequency.

Supersonic Oscillations. Some amplifiers oscillate at a frequency above 50,000 cps. A high setting of treble boost may start off the oscillation, in which case a reduction of the setting will usually stop the trouble. A change in the maximum possible setting of the treble control, a series resistor, or a shunt capacitor will eliminate the oscillation. These high-frequency waves are undamped and they load the amplifier so that all audio waves are badly distorted. The remedy may also be to add a 100- or 200-$\mu\mu$f

shunt capacitor to ground at an input-stage plate. The goal is to prevent the oscillation yet permit the stable amplification of audio frequencies. An oscilloscope at the output will detect the waves, showing amplitude as well as frequency. Note in Fig. 26 that R and C are shunted across the 47-K plate load resistor of the first tube; this discourages supersonic oscillations.

Ringing. This is a case which happens oftener than might be suspected. It can be detected by watching the output voltage on an oscilloscope, as steep wave fronts are applied to the input. A typical ringing picture is Fig. 77 when the input is a 10,000-cps saw-tooth. The steep wave front starts an amplifier to ringing, if there is any tendency toward instability. Figure 76 is the feedback circuit; across the 8-ohm output transformer secondary is 600 ohms plus 10 ohms to ground, with the input-tube cathode connected to the junction. The ringing frequency of about 90 kc is reduced by the addition of a mica capacitor C, until the saw-tooth wave is sharp with $C = 0.006$ μf as shown in Fig. 77.

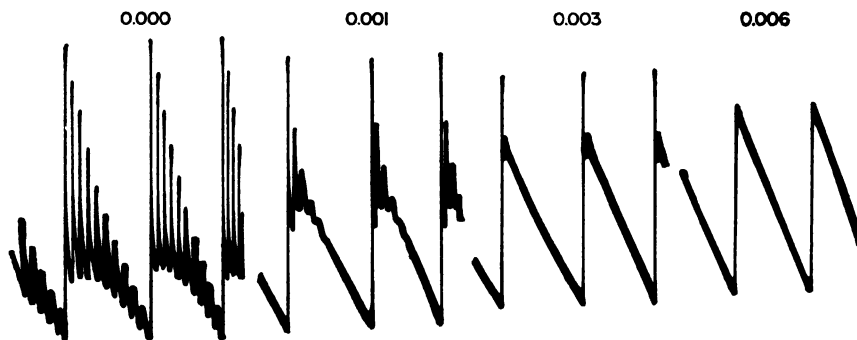



Fig. 77. Ringing pictures, using 10,000-cycle saw-tooth, as capacitor C across feedback resistor is changed.

Crossover Network. These may cause oscillation if the amplifier is not stable. One sure way to know is to connect all speakers and crossovers, supply a saw-tooth source, and watch the output voltage on an oscilloscope as the controls are all turned to their maximum positions.

Wrong Power-stage Feedback. This cannot exist without damage to an amplifier or the output transducer. It can happen only in a newly constructed amplifier. A loud squawk in the loud-speaker is evidence that the amplifier is acting as an oscillator because feedback is connected in reverse phase. Removing the feedback wire and reconnecting it in proper phase may mean either a change of a primary output transformer wire or an output transformer secondary wire, depending on how the feedback is arranged.

Too Much Feedback. This usually results in a low-frequency continuous oscillation. Remedy is to use less feedback, more resistance in the negative feedback loop from output to input. Positive and negative feedback have their places in special applications.[1] In general, negative voltage feedback is the rule,[2] although negative current feedback is sometimes employed (see Fig. 29).

Insufficient Decoupling between Stages. This produces a frequency so low that it may be inaudible, or it may be a "putt-putt" motorboat sound. This will show on an output voltmeter, and if the loud-speaker cone is not concealed, the cone can be seen to move in and out slowly. More series resistance and more shunt capacitance in the plate power supply are indicated. Two stages of RC-coupled tubes fed from a common positive source of voltage cannot regenerate. The normal result is more hum when the capacitance is inadequate. But for three stages, adequate decoupling is needed. A larger series resistor or a larger capacitor to the input-stage voltage supply is usually the answer.

[1] Childs, U. J., Loudspeaker Damping with Dynamic Negative Feedback, *Audio Eng.*, **36**, 11, February, 1952; Positive Current Feedback, *Audio Eng.*, **36**, 21, May, 1952. Clements, W., It's Positive Feedback, *Audio Eng.*, **36**, 20, May, 1952.

[2] Villchur, E. M., Negative Feedback, *Audio*, **39**, 22, June, 1955.

CHAPTER 16

ELECTRONIC OSCILLATORS

By J. K. Clapp, Frank D. Lewis, Robert A. Soderman, and Richard W. Frank[1]

Electronic oscillators are energy-conversion devices which draw electrical power from a d-c source and convert it into electrical oscillations. The oscillations so generated are alternating currents of a large variety of frequencies, wave forms, and power levels. In this chapter are discussed electronic oscillators generally, including oscillators using vacuum tubes, gas-filled tubes, and transistors as the energy converters. In addition, certain oscillators of the atomic and molecular type are included,

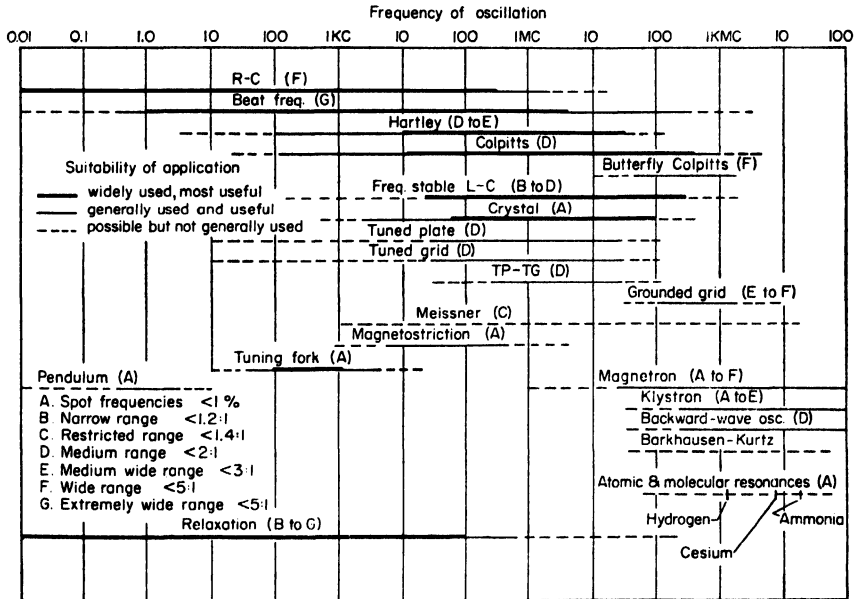



Fig. 1. Tuning ranges commonly afforded by adjustable tuning elements.

and some mechanical-electric systems having properties paralleling those of vacuum-tube oscillators.

Most electronic oscillators consist of a negative-resistance device and a control system including frequency-determining elements and amplitude-regulating elements. The negative-resistance device is usually an amplifier with the output coupled back to the input through the control system in the necessary manner to produce oscillations. The principal exceptions to this general rule are the atomic and molecular oscillators which may produce oscillations without this particular feedback circuit.

[1] General Radio Company.

The versatility of the electronic oscillator is evident in the wide frequency range covered and in the facility with which the wave forms desired can be produced. The most nearly constant frequencies are produced by electronic oscillators, and also frequencies which can be varied at will over wide ranges. Further discussion of the properties of electronic oscillators is given below, with references to the literature.

The chart of Fig. 1 lists the several types of oscillators discussed in this chapter arranged according to suitability of application and showing the general frequency coverage of each type.

VIBRATOR-CONTROLLED OSCILLATORS

In this section are grouped the oscillators capable of the greatest frequency stability. Each of these oscillator systems or circuits makes use of a vibrating system or device exhibiting a strong resonance at some particular frequency.

1. Atomic Resonance. The frequencies peculiar to the atoms of matter are thought to be fundamental constants of these atoms. In particular, the relationship $\epsilon = h\nu$ relates the energy ϵ involved in the change of the energy state of an atom when it

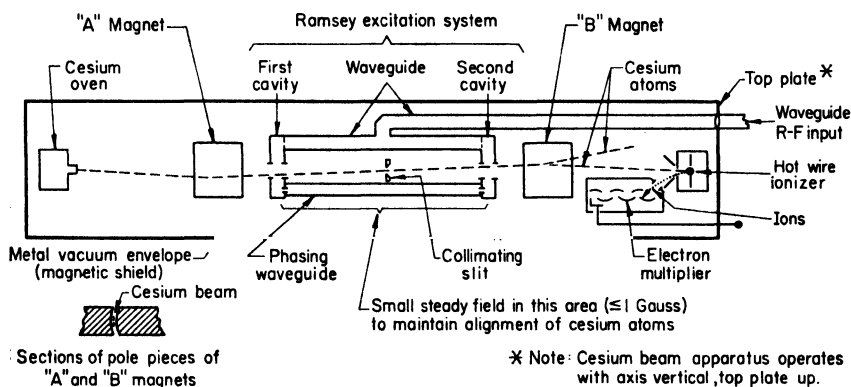



FIG. 2. Cesium atomic-beam frequency-standard apparatus.

absorbs or radiates a frequency ν. The symbol h represents Planck's constant (6.554×10^{-27} erg-sec). The absorption or radiation of energy by an atom is a quantum phenomenon, as there exists in each atom the possibility of only certain discrete states of energy. If r-f energy is coupled to atoms of a given substance at a frequency corresponding to one of the energy-level transitions permissible in those atoms, energy will be absorbed from the r-f field, resulting in the transition of some of the atoms to energy-level states higher than they occupied before the exposure to the r-f field. A number of methods have been devised for the recognition of such energy transitions in atoms.

One device which has been applied most successfully to the detection of atoms in various energy states is the *atomic-beam* apparatus. This apparatus using cesium atoms has been used as a frequency standard[1] (see Fig. 4).

The cesium-beam apparatus functions as follows: Cesium atoms are evaporated thermally from the source at the bottom and pass through the apparatus from bottom to top. The beam is operated vertically to avoid deflection of the atomic beam by gravity. The cesium atoms exist in two different energy states recognizable by the direction of deflection of the atoms as they pass through a magnetic-field gradient perpendicular to their line of travel. Atoms in one state will be deflected to the left by the magnetic gradient while those in the other state will be deflected to the right. Both magnets are oriented in the same way so that an atom remaining in the same energy state at the B magnet as it was at the A magnet

[1] Lewis, F. D., Frequency and Time Standards, *Proc. IRE*, **43**, 1046, September, 1955.

would receive two deflections in the same direction and miss the detector. However, if the atom changes its energy state between the two magnets, its deflection will be of opposite sign at the top magnet, and the atom will be deflected so that it hits the detector. A cesium atom can be made to change its energy state by adding or subtracting a quantum of energy of the correct value. This can be accomplished by passing the atom through a magnetic field oscillating at exactly the correct frequency, which is approximately 9,192.387 Mc for cesium. The cesium-beam apparatus shown uses a two-cavity, or Ramsey, excitation system which increases the resolution of the over-all system with respect to frequency. The atoms which have changed state are received at the detector, which, in one form, consists of a hot-wire ionizer and electron multiplier of the secondary-emission type. The number of atoms influenced by the r-f field is a function of the oscillation frequency, the number being greatest when the frequency is exactly equal to the required frequency for the quantum of energy needed to make the atom undergo a transition from one energy state to the other. The use of the divided excitation system produces an interference effect in the absorption of energy by the atoms, evidenced by a sharpening

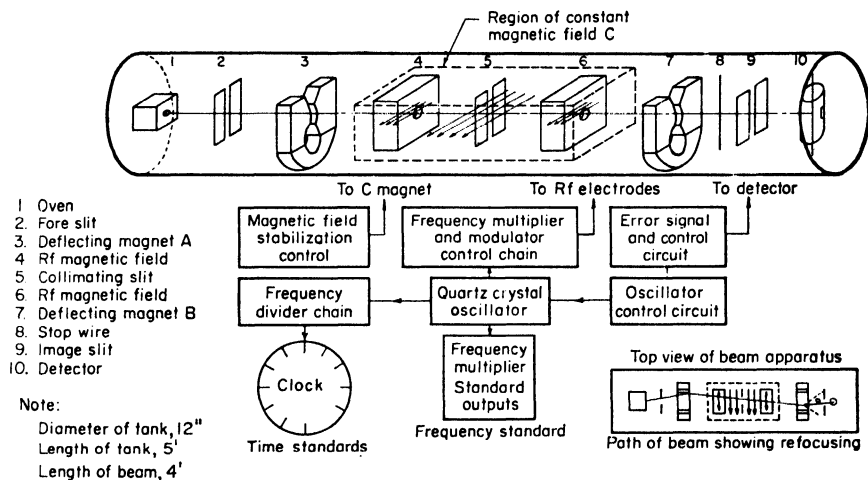



FIG. 3. Schematic of atomic-beam clock and frequency standard developed at National Bureau of Standards Boulder Laboratories.

of the main peak of detector current accompanied by the appearance of "side lobes" as the frequency of excitation is varied slowly about the correct value.

It is possible to apply servosystem techniques to the cesium-beam apparatus to provide automatic frequency control of the r-f source, and thus develop a frequency standard from the atomic cesium resonance. One such system makes use of a crystal-controlled oscillator in the h-f band followed by a multiplier and a frequency synthesizer to produce the nominally correct frequency at the 9,192-Mc level. Inserted in the multiplier chain is a phase modulator, which, when followed by a large multiplication ratio in the multiplier chain, produces an f-m signal sweeping across the cesium resonant frequency. If the deviation is symmetrical about the resonant frequency, the detector current will contain only second-harmonic components of the modulation-ratio frequency, but if the deviation includes a fixed-error term, a fundamental-frequency component will appear. The phase and magnitude of the fundamental modulation-frequency component in the detector current are used to control a servosystem to maintain the r-f source on the correct frequency.

Other atomic resonances are thought to be suitable for similar applications, notably the resonances of those atoms having, like cesium, only one electron in the outermost shell. The great interest in atomic and molecular resonances stems from the belief that they represent constant-frequency references unaffected by astronomical phenomena.

2. Molecular Resonance. Since the variety of molecules is infinitely greater than that of atoms, one might be led to speculate that it would be easy to find a molecule with a convenient quantum absorption frequency and other properties making it possible to use such a molecule as a resonator for an electronic oscillator. Although much is now known[1] of the microwave spectroscopic properties of molecules, the choice of a suitable molecular resonance is difficult and the application of such a resonance is subject to many difficulties.

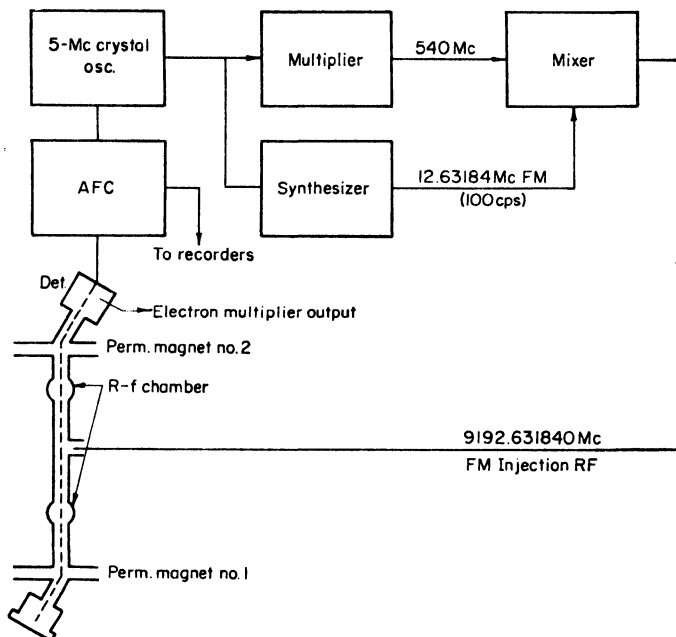



FIG. 4. Block diagram of National Company Atomichron cesium oscillator.

The most attractive molecule appears to be ammonia, from a microwave-resonance standpoint. Much fundamental work on microwave absorption has been done on ammonia,[2] and several methods have been devised for using the ammonia resonance at approximately 23,870 Mc for frequency standardization.

The first method to be applied practically[3] employed ammonia absorption in a wave-guide container. The r-f energy traversing an ammonia-filled wave guide was frequency-modulated so that it swept across the absorption frequency of the ammonia. As the ammonia absorbed energy, it reduced the energy received at the receiving end of the wave guide, and an absorption "pip" was produced which could be used to aid in adjustment of the oscillator frequency to keep the oscillator centered on the ammonia absorption. Fairly extensive investigation of this method for using ammonia has been carried out. Practical difficulties have restricted the stability of such systems and have led to continued

[1] Strandberg, M. W. P., "Microwave Spectroscopy," Methuen, 1954. Gordy, W., W. V. Smith, and R. Trambarulo, "Microwave Spectroscopy," Wiley, 1953.

[2] Townes, C. H., The Ammonia Spectrum and Line Shapes near 1.25 Cm Wavelength, *Phys. Rev.*, **70**, 665, November, 1946. Good, W. E., The Inversion Spectrum of Ammonia, *Phys. Rev.*, **69**, 539, May, 1946. Bleaney, B., and R. P. Penrose, Ammonia Spectrum in the 1 Cm Wavelength Region, *Nature*, **157**, 339, May, 1946.

[3] Hershberger, W. D., and L. E. Norton, Frequency Stabilization with Microwave Spectral Lines, *RCA Rev.*, **9**, 38, March, 1948. Lyons, H., The Atomic Clock, an Atomic Standard of Frequency and Time, *Nat. Bur. Standards Tech. News Bull.*, **33**, 17, February, 1949.

research in attempts to overcome the inherent difficulties of this system. The difficulties arise from Doppler effect and collision broadening of the absorption line and signal-to-noise ratio troubles which are linked to saturation effects.

A simpler method[1] using an ammonia-filled resonator in a Pound-type stabilizer system has been used, but it also is subject to the inherent difficulties of the absorption scheme.

A relatively recent development has enabled the use of the ammonia gas itself to produce amplification by what is known as stimulated emission of radiation. Devices of this sort are called *Masers*[2] (from the initial letters of *M*icrowave *A*mplifier by *S*timulated *E*mission of *R*adiation). An oscillator making use of this principle (see Fig. 5) consists of a source of ammonia which provides a stream of gas molecules traveling through the focusing electrodes

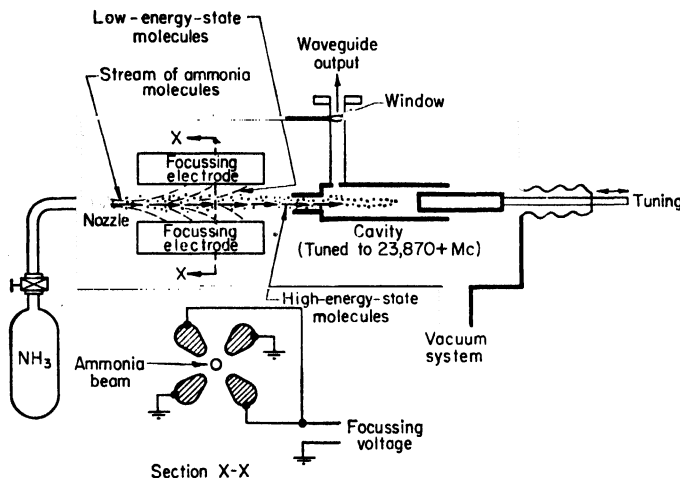



FIG. 5. Ammonia oscillator of Townes.

into a cavity resonator tuned to the ammonia frequency. As the molecules go through the focuser, those in the lower energy state are diverged and those in the upper state are converged and focused at the resonator. When the molecules go inside the resonator, they are acted upon by the electric field in the resonator, since the ammonia molecule in the upper (inversion) energy state has an external electric dipole moment. Noise energy in the resonator thus triggers the ammonia molecules to cause them to emit quanta to the field in the resonator, the molecules then undergoing transition to the lower energy state by the emission of the quanta. Since the molecules themselves emit the power, and since the trigger mechanism is linked directly to the phase of the r-f field acting on the molecules, the process produces a phase-coherent field in the cavity. The amount of power produced by an oscillator of this sort is very small, being estimated at 10^{-9} watt. The frequency stability is relatively high. It is expected that this type of ammonia oscillator will become useful as a frequency standard.

Although most of the work on molecular resonances for frequency control has been done on ammonia, other molecules may prove useful in this respect.[3] Detection of

[1] Smith, W. V., J. L. G. de Quevedo, R. L. Carter, and W. S. Bennett, Frequency Stabilization of Microwave Oscillators by Spectrum Lines, *J. Applied Phys.*, **18**, 1112, December, 1947.

[2] Gordon, J. P., H. J. Zeiger, and C. H. Townes, Molecular Microwave Oscillator and New Hyperfine Structure in the Microwave Spectrum of NH_3, *Phys. Rev.*, **95**, 282, July 1, 1954; The Maser—New Type of Microwave Frequency Standard, and Spectrometer, *Phys. Rev.*, **99** (4), 1264, Aug. 15, 1955. Shimoda, K., T. C. Wang, and C. H. Townes, Further Aspects of the Theory of the Maser, *Phys. Rev.*, **102** (5), 1308, June 1, 1956.

[3] Bloembergen, N., Proposal for a New Type Solid State Maser, *Phys. Rev.*, **104** (2), 324–327, Oct. 15, 1956. See also *Bell Lab. Record*, **35**, 109, March, 1957.

hydrogen radiation from extraterrestrial sources has spurred investigation of microwave-energy transitions in gases of many different varieties, and may lead to discovery of further useful microwave-energy-transition levels of this sort.

CRYSTAL-CONTROLLED OSCILLATORS

3. Piezoelectric Crystals. No attempt will be made here to cover, even briefly, the very broad fields of crystallography and piezoelectric crystals. A comprehensive treatise is W. G. Cady's book "Piezoelectricity."[1] References on the processing and applications of quartz crystals which are very useful are those of Mason,[2] Heising,[3] Vigoreux,[4] D'heedene.[5]

Because of its availability, suitable mechanical properties, and reasonable cost, quartz is utilized almost entirely in producing crystals for frequency-control purposes.

The natural crystals rarely show a complete crystal form, which in quartz is a hexagonal prism with pyramidal ends. If the cap faces are present, visual inspection discloses whether the crystal is "right-" or "left-hand," meaning the sense in which the crystal rotates the plane of polarization of light passed through the crystal in the direction of the optic axis. If the crystal is broken, or worn, as in "river" quartz, elaborate testing is required to establish the handedness and the directions of the axes.

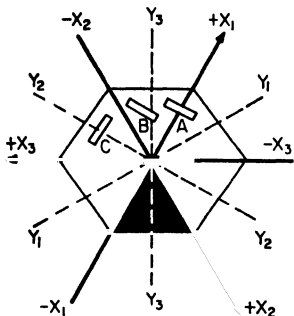



Fig. 6. Illustrating that the "axes" are directions in quartz.

The "axes" in crystals are not components of a fixed reference frame but are *directions* in the quartz. In Fig. 6, at any point on the crystal section, the direction perpendicular to the surface of the section (*i.e.*, perpendicular to the plane of the paper) is the "optic axis." At such a point, the direction parallel to X_1 is an electric axis, and so on.

A plate cut with its faces perpendicular to the direction X_1 would be an "X-cut" plate, no matter where located in the section, as at A, B, etc.

If such a plate is etched, then observed in a microscope, the surface toward the arrow of X_1 shows one kind of pit pattern, while the surface away from the arrow of X_1 shows an entirely different pit pattern. This is a visual demonstration of the fact that the crystal shows a polarity and that one end of the electric axis is positive while the other is negative.

By virtue of the "piezoelectric effect," if a quartz crystal is put in compression or in tension, electric charges will be developed on certain faces. Conversely, if these faces are charged, then an elongation or contraction of the quartz will result. By use of a vacuum-tube amplifier, a voltage developed on the electrodes by vibration of the plate is amplified and fed back in step so as to maintain the crystal vibration. The coupling between the mechanical system of the vibrator and the electrical drive system is extremely small. Even so, care must be taken to limit the drive on the crystal lest the crystal be fractured because of the extremely small internal losses in quartz. The equivalent electric circuit of a crystal was described by Van Dyke[6] and Mason.[7] In Fig. 7 the branch $L_1C_1R_1$ represents the motional impedance of the vibrator, where L_1 is proportional to the mass, C_1 to the elastance, and R_1 to the dissipation. C_0 represents the static capacitance of the quartz between the electrodes.

[1] Cady, W. A., "Piezoelectricity," McGraw-Hill, 1946.

[2] Mason, W. P., "Piezoelectric Crystals and Their Application to Ultrasonics," Van Nostrand, 1950; "Electromechanical Transducers and Wave Filters," Van Nostrand, 1942.

[3] Heising, R. A., "Quartz Crystals for Electrical Circuits," Van Nostrand, 1946.

[4] Vigoreux, P., and C. F. Booth, "Quartz Vibrators and Their Applications," H. M. Stationery Office, London, 1950.

[5] D'heedene, A. R., "Quartz Crystals for Electric Circuits," Chap. 14, p. 456, Van Nostrand, 1946.

[6] Van Dyke, K. S., *Proc. IRE*, **16**, 742, 1928.

[7] Mason, W. P., *Proc. IRE*, **23**, 1252, 1935.

The ratio C_1/C_0 represents the ratio of energy stored in the mechanical to that in the electrical system.

The simplest cuts are the X-cut and Y-cut plates as at A and C, respectively, in Fig. 6. The X-cut plate vibrates in a *thickness extensional mode*, where the large surfaces of the plate alternately move apart or come together. The Y-cut plate vibrates in a *thickness shear mode*, where the upper surface alternately slides one way, then the other, while the lower surface moves similarly but in the opposite direction.

Fig. 7. Quartz plate equivalent circuit.

4. Temperature Coefficient. X-cut plates have temperature coefficients of frequency which lie in the range of -5 to -25 ppm per °C, while the Y cuts range from $+25$ to $+100$ ppm. Making a cut at some angle between the two will not yield a zero temperature coefficient because of the different modes of operation.

Development of low-temperature-coefficient quartz crystals has extended over the whole period of use of quartz plates. The history and the properties of the various cuts are well covered in Cady's book.[1] Basic coverage on low-temperature-coefficient crystals is given by Lack, Willard, and Fair.[2]

As a matter of fact, only one crystal cut can in any truth be claimed as a "zero-temperature-coefficient" cut, that is, the GT cut.[3] When properly made it has

Fig. 8. Temperature dependence of various crystal cuts. (*From Bell System Tech. J.*)

a very small temperature coefficient over a wide temperature range, 0 to 100°C. All other "zero-temperature-coefficient" crystals have parabolic characteristics such that, at the "turn-over" point, the slope of the frequency-temperature curve is zero. At this point, no change in frequency occurs over very small changes in temperature. In other words, the crystal has a zero temperature coefficient of frequency *at a single temperature only*, as is easily seen in Fig. 8.

The commonly used quartz plates, having zero temperature coefficient of frequency at a single temperature, are "rotated Y-cut" plates, where one edge is in the direction of the X axis, and the other, instead of being in the direction of the Z, or optic, axis,

[1] Cady, *loc. cit.*

[2] Lack, F. R., G. W. Willard, and I. E. Fair, Some Improvements in Quartz Crystal Circuit Elements, *Bell System Tech. J.*, **13**, 453, July, 1934.

[3] Mason, W. P., A New Quartz Crystal Plate, Designated the GT, Which Produces a Very Constant Frequency over a Wide Temperature Range, *Proc. IRE*, **28**, 220, May, 1940.

makes an angle with it. There are two angles, one positive and one negative, where zero-temperature-coefficient plates result. It is interesting to note that these angles are related to the minor rhombohedral faces of the crystal pyramidal cap, and that the plates are cut approximately parallel with these faces, as can be seen from Fig. 9.[1]

The temperature at which the turn-over point occurs is a very critical function of the angle at which the plate is tipped away from the Z axis. If uniform performance

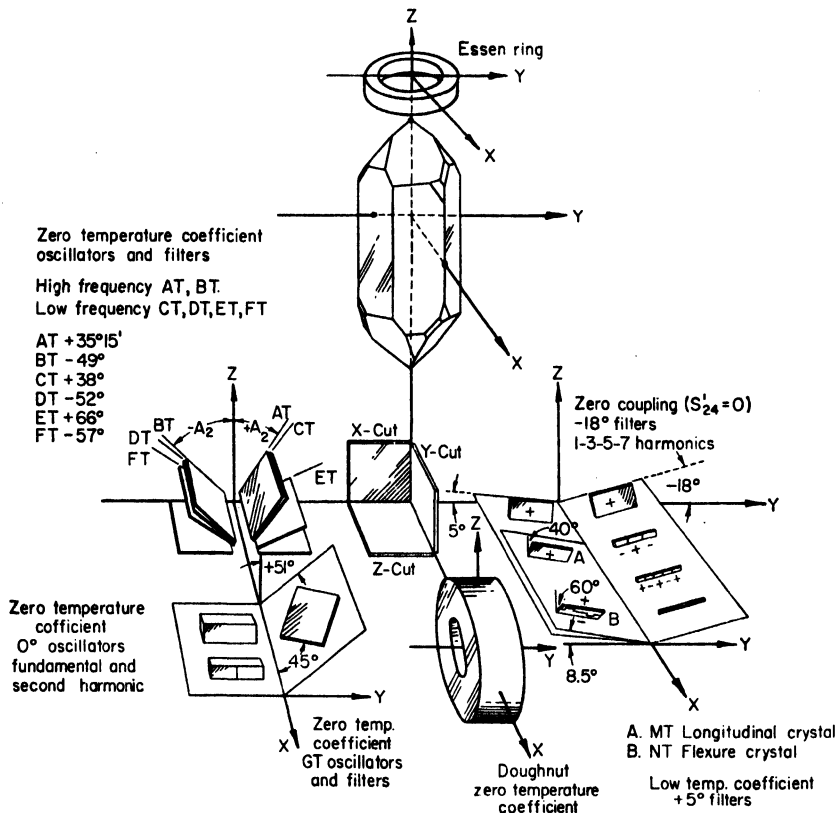



Fig. 9. Types of crystal cuts. (*Modified from Heising, "Quartz Crystals for Electrical Circuits."*)

of these crystals is required, continuous checking of the angle of tilt must be carried out during processing. It is not sufficient to test the angle of a blank, for example, and assume that the angle will be maintained as the crystal is ground to frequency.

The thickness-shear mode AT and BT crystals are widely used for frequencies above 500 kc. These crystals can be used in overtone modes giving substantially higher frequencies—up to 250 Mc or more.

5. Overtone Operation. The mechanism of overtone operation is illustrated in Fig. 10. In *a* the fundamental mode is shown, where the displacement resulting from charging the electrodes is shown. If a suitable circuit is used, the thickness of the crystal can be made to vibrate in several sections, as in Fig. 10, where five

[1] Koga, Isaac, *Proc. IRE*, **24**, 510, March, 1936.

sections are illustrated, each of one-fifth of the thickness, and consequently the frequency will be five times as high. The overtone frequencies are not *exact* multiples of the fundamental frequency but are very nearly so.

It is evident from the figure that the electrodes are such that the polarity of the crystal slabs must always agree. If an even number of slabs existed, the top and bottom faces would be of like sign and no excitation would be possible. Consequently only odd overtones can be excited.

The shunt capacitance of the quartz imposes a limitation on how high an overtone can be used successfully in an oscillator requiring the crystal to present an inductive reactance. By means of a special holder, which in effect provided a capacitance bridge to neutralize the shunt capacitance, overtones as high as the twenty-third have been used. A parallel-tuned circuit placed across the crystal and tuned to the desired overtone frequency accomplishes a similar purpose.

FIG. 10. Fundamental and overtone operation.

6. Crystal Mounting. The mounting of a quartz plate is most important in determining its performance. Quartz plates for the most accurate requirements are now made with lens-shaped instead of flat surfaces. The central area may or may not be flat and is the area where the electrodes are plated. The edges of the crystal, if it is of proper size for the frequency required, are nearly quiescent so that attachment of the mounting and connecting wires does not appreciably disturb the vibration. As there is no motion at the edges, no energy from the vibrating portion is lost through the supports.

The crystal mount is placed in an evacuated glass bulb which not only eliminates losses caused by ultrasonic radiation into the air, but also prevents air loading, contamination, moisture, etc., from affecting the crystal.[1]

An extensive study of quartz plates by Bömmel, Mason, and Warner[2] answers many questions which heretofore have apparently had no or, at most, ambiguous answers. The performance of plates in air (or inert gases) differs considerably from that in vacuum. Conclusions arrived at in the first case do not apply in the second.

For 15-mm-diameter contoured quartz plates in vacuum, the following may be concluded:

1. Q is greater by three times than in air.
2. Q is no longer a function of the overtone (or thickness) of the plate. Observations on 10-Mc plates at the third, fifth, and ninth overtone, giving 3:1 difference in thickness and 27:1 difference in equivalent L and R, showed the same Q at all overtones.
3. Q does not change with plating thickness in a 3:1 range.
4. Carefully polished surfaces yield Qs no more than 10 per cent above those obtained with carefully lapped and etched surfaces.
5. Above 30 Mc, evacuation produces no measurable increase in Q.
6. At 100 Mc, the best Q is about 150,000.

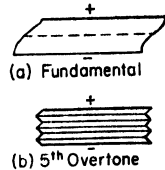
Low-temperature Quartz Crystals. The internal loss $1/Q$ increases as the temperature is reduced down to 52°K, where a sharp internal friction peak occurs. The loss then decreases, rises again to a broad lower peak around 21°K, then decreases to a value of 2×10^{-7} at 6°K and remains constant down to 1.5°K (lowest temperature reached). The constant loss has been ascribed to mounting loss. The indicated relaxation frequency at room temperature is about 160 Mc.

This agrees with experience that the Q of quartz plates decreases as the frequency increases. Beyond 160 Mc, the Q should begin to increase as the frequency is increased.[3]

If crystals are operated at liquid-hydrogen temperature (4.3°K) the Q of an AT-cut 5-Mc crystal is at least twice that at room temperature and aging is greatly reduced. The temperature coefficient of frequency is $0.01/10^6/°K$. Indicated stabilities of 1 in 10^{11} are possible.

[1] Warner, A. W., High-frequency Crystal Units for Primary Frequency Standards, *Proc. IRE*, **40**, 1030, September, 1952; also *Proc. IRE*, **42**, 1452, September, 1954.

[2] Bömmel, H. E., W. P. Mason, and A. W. Warner, Dislocations, Relaxations and Anelasticity of Crystal Quartz, *Phys. Rev.*, **102** (1), 64–71, 1956.

[3] George, W. D., Frequency Standard at Low Temperature, *Proc. 10th Annual Symposium on Frequency Control*, May, 1956, p. 197. Searles, C. E., R. A. Sykes, and I. E. Fair, "Information Bulletin on Quartz Crystals," Armed Services Electro Standards Agency, Fort Monmouth, N.J.

If the decrease of aging follows the increase in time of chemical reactions, the aging of a quartz plate would be reduced by a factor of the order of 10^{10}. The aging in 40 years at low temperatures would be less than that in 1 sec at 50°C.

The writers[1] conclude that the best frequency for frequency-control purposes is a function of the size of quartz plate available. For 15-mm-diameter plates, the best frequency is approximately 5 Mc. Tests with 30- and 90-mm-diameter plates show the Q at 1 Mc is not degraded by loss of energy at the edge of the plate in these larger sizes. Maximum Q is about 10×10^6.

7. Essen Ring. A form of quartz crystal for low-frequency precise standards giving excellent performance is the Essen ring.[2] The ring is 90 mm in diameter and vibrates in an extensional mode of six sections. The ring is supported on threads touching the ring at points 120 deg apart.

The frequency of oscillation is a function of the mean diameter, which probably accounts for the very small aging observed. If material is lost from the surfaces, the loss on the outside is nearly balanced by the loss on the inside, so that the frequency change is very substantially less than in more usual types.

The crystal has a temperature-frequency curve similar to that of bars, and by choice of the dimensions of the cross section the temperature of the turn-over point can be controlled.

8. Servocontrolled Piezoelectric Oscillators. A crystal oscillator circuit providing unusual means for ensuring that the crystal is the principal frequency-controlling element is the servocontrolled oscillator of Lea,[3] of which a variation is that of Sulzer.[4] In Fig. 11, a Gouriet-Clapp oscillator is shown modified to include the servocontrol circuit. A bridge is constructed by adding a resistor R_1 in series with the crystal and adding two more arms of R_2 and R_3 and X, with R_2 and $R_3 \gg R_0$ to minimize shunting effects on the crystal branch. The reactances X are arranged to be differentially modulated by a modulating device which is a part of the servosystem. By arranging for the output of the a-m detector on the bridge to be compared with the modulation amplitude and phase of the modulating device, the sense and magnitude of the bridge unbalance are measured and used to control a servodrive which readjusts the oscillator circuit to restore the frequency for which the bridge has no reactance unbalance. The bridge is set up to balance initially for the resistance values chosen. Hence the bridge balance is a function of the frequency of the driving source only, balance being obtained at series resonance of the crystal. This system provides

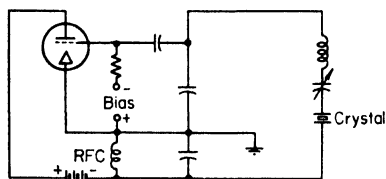



FIG. 11. Servocontrolled oscillator, a modification of the Gouriet-Clapp circuit.

[1] Bömmel, Mason, and Warner, *loc. cit.*

[2] Essen, L., A New Form of Frequency and Time Standard, *Proc. Phys. Soc.* (*London*), **50,** 413, 1938. Mitchell, H. T., and A. L. Dobbie, 100 kc/s Oscillator of High Precision Incorporating an Essen-type Quartz Ring, Paper presented at Congrès International de Chronometrie, Paris, France, October, 1954.

[3] Lea, N., Quartz Resonator Servo—A New Frequency Standard, *Marconi Rev.*, **17,** 65, 3d quarter, 1954.

[4] Sulzer, P. G., High-stability Bridge-balancing Oscillator, *Proc. IRE*, **43** (6), 701, June, 1955.

correction for any variables affecting the oscillato frequency, which is the frequency of the signal applied to the crystal bridge circuit, in so far as the servo ystem can correct for these variables. There is always a small residual error in a servosystem, but by the provision of a large amount of gain in the detector device, a large amount of correction can be obtained. One of the important properties of this circuit is that the added stabilization system does not work at the crystal frequency but at a fixed frequency in the a-f range (the modulation-rate frequency) and hence is useful for stabilizing crystal oscillators in the h-f and v-h-f regions as well as in the m-f and l-f bands where conventional negative-feedback stabilization of amplifiers is most effective. Also, it is possible to correct for such items as cathode-interface impedance by this method.

Lea applied this circuit to a 5-Mc oscillator using a rotating-toothed-wheel inductance modulator for reactance variation and obtained stabilities of 1 part in 10^{10} or better when tubes were changed and supply voltages varied over wide ranges. The crystal frequency for series resonance is thus used as the reference for the servosystem to provide improvements of 10^4 or 10^5 in over-all long-term stability. The short-term, or phase, stability of the oscillator is determined by the crystal, the effective Q of

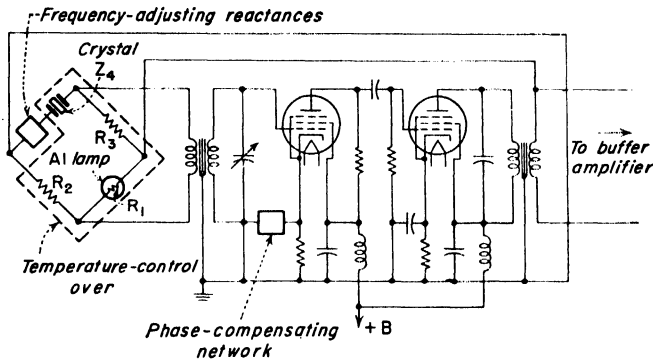



FIG. 12. Meacham's bridge-stabilized oscillator.

which has been degraded to one-half its original value by the addition of the resistor R in series with the crystal.

9. Conventional Piezoelectric Oscillators. In Lea's oscillator, described above, the only variation in frequency which can occur is that caused by changes in the constants of the reference crystal itself. Such changes would result in a change in the frequency at which the reference phase occurs.

In conventional circuits, changes in frequency result from any changes in the constants of the crystal but also from any changes in circuit reactances. Changes in the constants of the vibrator can result from the gradual relief of strains in the quartz, exposure to contaminating influences, chemical or physical changes in the plating on the surfaces, migrations of metals involved in the plating or soldering of supports, changes in the supports, and loss of material from the surface of the quartz.

Changes in circuit reactances can result from aging of vacuum tubes, development of cathode-interface impedance, changes in electrode voltages, changes in tube structure with temperature or time, changes in operating level, and a variety of other factors.

10. Driving Circuits. The Meacham[1] oscillator in Fig. 12 represents a high degree of development in making the oscillator frequency nearly independent of the driving circuit. The frequency variations remaining, though small, are none the less troublesome for the most precise requirements, and considerable effort has been expended in still further reducing them.

[1] Meacham, L. A., The Bridge-stabilized Oscillator, *Proc. IRE*, **26** (10), 1278, October, 1938. Also *Bell System Tech. J.*, **17** (4), 574.

One approach was that of Lea, whose oscillator has been described above. The U.S. Bureau of Standards eliminated the driving circuit entirely, using the quartz crystals as resonators only. This is satisfactory as a reference standard but does not yield a working frequency standard.

In spite of their defects, conventional crystal oscillators are still our working tools. A thorough understanding of the operation of the circuits will, in time, lead to still further improvements in performance.

11. Meacham Oscillator. Fundamentally, this is a tuned pentode amplifier with the output stepped down to a low-impedance bridge by the tuned output transformer. The bridge contains the quartz crystal and a small lamp, which is the means for regulating the operating level, in two of the arms. The other two arms are resistors. The bridge output is stepped up through a tuned input transformer to the grid of the pentode amplifier.

The bridge arms are arranged so that, when oscillation starts, the lamp filament is cold, with the lamp resistance at a low value, and the bridge is unbalanced in such a way as to produce a large output voltage from the bridge in a sense to give positive feedback. The amplitude of oscillation builds up rapidly and the lamp-filament temperature rises quickly. As the lamp resistance increases, the bridge approaches, but never quite reaches, balance. The bridge output voltage is thus reduced to a very small value, which stabilizes at the level where the loss through the bridge is just equal to the gain through the amplifier back to the bridge input. Large changes in bridge loss result in minute changes in input level; the bridge therefore is a most effective automatic level control.

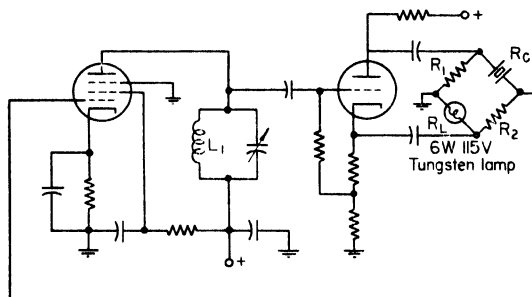



FIG. 13. Modified Meacham oscillator.

The arms of the bridge are intended to be nonreactive. This means that the crystal must operate at zero reactance, which, for low-frequency crystals such as those commonly used in primary frequency standards, is very nearly at series resonance.

Meacham gave an analysis of the operation of the bridge oscillator in which considerable emphasis was placed on the effective cancellation of the resonator resistance by the balancing of the bridge. This results in an enhanced slope of the phase characteristic of the bridge output, which led to the popular statement that "the bridge multiplied the Q of the crystal."

The use of two tuned transformers makes Meacham's circuit difficult to adjust and more sensitive to variation in circuit parameters than it need be. For greater simplicity in operating with crystals of different frequencies, a circuit was proposed[1] using only one tuned circuit (Fig. 13). In adjusting this circuit, if means are available for phase indication, the tuning can be set easily and accurately by phase rather than by amplitude indications.

Post and Pit[2] point out that other viewpoints are equally valid. They also contribute a most important factor in analysis of the oscillator. This is that, while attention has been almost exclusively directed (most especially in the United States) to improvement of the quartz resonator to obtain higher Q, equally important gains

[1] Clapp, J. K., An Analysis of the Bridge-controlled Oscillator, *Gen. Radio Experimenter*, May, 1944.

[2] Post, E. J., and H. F. Pit, Alternate Ways in the Analysis of the Feedback Oscillator and Its Application, *Proc. IRE*, **39** (2), 169, February, 1951. Cady, *op. cit.*

are possible by improvement of the driving amplifier circuit. The stability of the oscillator depends on the *ratio* of resonator phase slope to amplifier phase slope, and making the amplifier phase slope *effectively zero over the very limited required frequency range* is just as important as obtaining the *greatest possible phase slope in the resonator.*

12. One-tube Oscillators. A general-purpose and simple piezoelectric oscillator is a one-tube circuit which is the basis of the so-called "series-tuned" circuits. A diagram is given in Fig. 14*a*. The reactances of C_1 and C_2 are made as low as possible

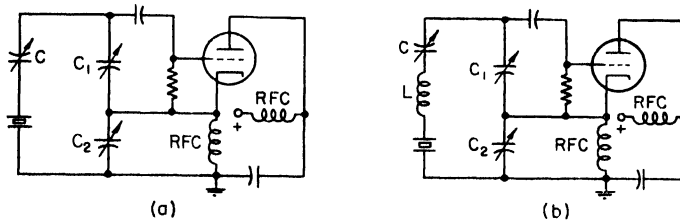



Fig. 14. General-purpose crystal oscillators. (*a*) Series-resonant crystal; (*b*) crystal adjusted to work with specified value of series capacitor *C*.

while still permitting oscillation. If frequency adjustment is desired, a reactance can be placed in series with the crystal, usually a small capacitance. The crystal must then be ground so its natural frequency is slightly low, and the value of series-capacitance must be specified. Several values of capacitance are used as "standard," 32 $\mu\mu$f being used most frequently.

A better system is to use an inductive and capacitative reactance in series, of such values that, if the crystal is replaced by a resistor, the frequency of oscillation is the same as the crystal frequency. The crystals would then be ground as "series-resonant" (Fig. 14*b*).

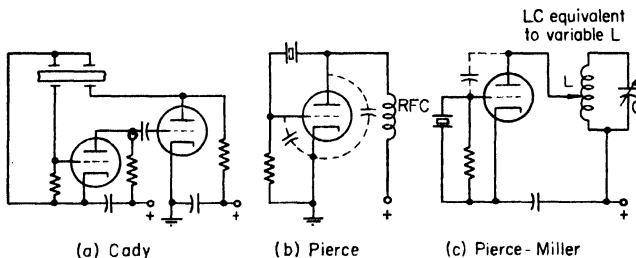



Fig. 15. Crystal oscillator circuits.

The older piezoelectric oscillators are little used today. The early forms developed by Cady, Pierce, and Miller are shown in Fig. 15, and are described by Cady.[1]

In *Cady's circuit* the crystal acted as a coupler between output and input of the amplifier, where a voltage appeared at the grid as a result of the mechanical vibration produced by the electrodes connected to the plate circuit. Variations of this type of crystal are now used in filters rather than in oscillators.[2]

In *Pierce's circuit*, the crystal is connected between grid and plate, and a resistor or choke coil is utilized in the plate circuit, later replaced by a tuned circuit. This is a *Colpitts circuit* with the tube capacitances acting as the circuit capacitances and the crystal operating above its resonant frequency to provide an inductive reactance.

[1] Cady, *op. cit.*, pp. 286, 494.

[2] Post, E. J., and J. W. A. Van der Scherr, Bridge Stabilized Oscillators and Their Derivatives, *Het P. T. T.*, Bedrijf, **6** (4), 86, September, 1944.

In the *Pierce-Miller circuit*, the crystal is connected between grid and cathode, and a tuned circuit is used in the plate circuit with the plate tapped down to improve the power output. Here the grid and plate impedances are inductive, the crystal operating above resonance, and the grid-plate tube capacitance completes the oscillating (Hartley) circuit.

Neither the Pierce nor Pierce-Miller circuits can operate at the resonant frequency of the crystal, and the nearer to the resonant frequency they are adjusted, the more unstable and critical they become.

The "series-tuned" type of Fig. 14 is much to be preferred.

Phase-locked quartz-crystal oscillators can be used to great advantage in frequency multipliers.

OTHER VIBRATOR-CONTROLLED OSCILLATORS

13. Tuning-fork Oscillators. Vacuum-tube-driven tuning forks are useful for frequencies from 50 cycles up to a few thousand cycles. The principal advance over the years has been in the very substantial decrease in the temperature coefficient of frequency of the fork. The design first utilized a bimetallic fork where the heel of the fork was of one alloy and the end portions of the tines were of another alloy.

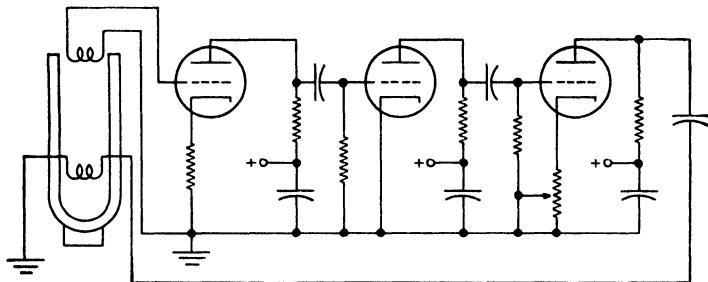



FIG. 16. Tuning-fork oscillator.

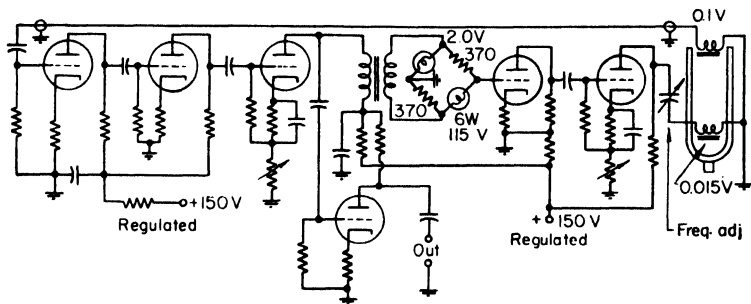



FIG. 17. Amplitude-regulated fork oscillator.

Later developments brought out the "sandwich fork" where the basic fork was of a uniform material. To correct the temperature coefficient, a thin slice of a different alloy was bonded on one side of the fork. This gives a structure which can be *adjusted* for temperature coefficient by grinding either the basic fork or the sandwich layer. Temperature coefficients of frequency less than 1 ppm per °C can be obtained in production.

The drive of precision forks is generally by electromagnetic excitation and pickup, as indicated in Fig. 16. Most fork oscillators have no provision for amplitude control, other than amplifier limiting. Improved performance can be obtained by use of

amplitude control, shown in Fig. 17. Fine adjustment of frequency can be made by phase control in the oscillator circuit.

14. Magnetostriction Oscillators. Of late years not much attention has been given to magnetostriction resonators, principally because of the comparatively high temperature coefficient of frequency. The rod alloys have a bearing on this coefficient, and searching for "zero temperature coefficient" is most disappointing because just as one creeps up on that most desirable state, the activity reaches zero first. Nature can be so contrary! A comprehensive paper was given by Pierce[1] discussing oscillator circuits and various magnetostrictive alloys.

The circuit for a magnetostriction oscillator is very simple, the principal variation being that the coupling is reversed in comparison with the usual oscillator (Fig. 18).

FIG. 18. Magnetostriction oscillator.

The frequency range of magnetostriction oscillators is from several hundred cycles to several tens of kilocycles.

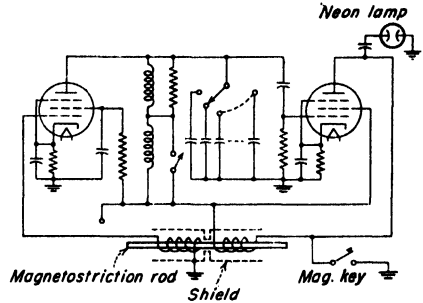
15. Pendulum Oscillators. Oscillators operating at very low frequencies can be stabilized by the use of pendulum resonators. The most highly developed forms of precision pendulum clocks[2] make use of electric impulses to drive a free pendulum in an evacuated chamber, related electric impulses being used to drive a slave clock with indicating mechanism. When the pendulum swings at a very small and constant amplitude in an evacuated chamber, the oscillation time is approximately $T = \sqrt{l/g}\,(\sin \theta)/\theta$. Enough energy is added on each swing, at the point of maximum velocity, to make up for loss in the supports. Constant-period oscillations result from this method of operation.

The Shortt and Riefler clocks which have been used as observatory time standards are made in the manner described above. Stabilities of the order of 1 part in 10^7 were not unusual for these clocks although they were affected by changes in the gravitational constant and by vibration of their mountings. Extreme care is necessary in the installation of such clocks to provide solid vibration-free installations in constant-temperature surroundings.

The contacts used on the clocks for controlling the impulse drives were mechanical in nature. Other means can be used such as photoelectric devices, or inductive or capacitive pickup systems. These systems would require the use of amplifiers to ensure proper functioning.

LC OR RESONANT-CIRCUIT CONTROLLED OSCILLATORS

16. Frequency-stable Oscillators. Here are emphasized those circuits intended particularly for uses where frequency variations caused by the driving tube are to be reduced to a minimum. Typical applications are as master oscillators in transmitters, heterodyne-frequency meters, beat-frequency oscillators, and finely and accurately adjustable frequency sources.

Over a period of two decades, workers in various parts of the world have attacked this problem; all have obtained the same basic result. To keep discussion within reasonable limits, assume the following restrictions:

[1] Pierce, G. W., Magneto-striction Oscillators, *Proc. Am. Acad. Arts Sci.*, **63**, 1–47, 1928. Also *Proc. IRE*, **17**, 42, 1929.

[2] Marrison, W. A., The Evolution of the Quartz Crystal Clock, *Bell System Tech. J.*, **27**, 510, July, 1948. Also published as "Bell Telephone System Monograph B-1593," Bell Telephone Laboratories, and in *Horological J.*, **90**, 274, May–October, 1948.

1. Frequency stability shall apply only to those effects produced by the driving-tube amplifier, due to (1) changes in tubes, (2) changes of electrode voltages, (3) mechanical changes with time, (4) temperature, or (5) changes in operating level.

2. The changes in the frequency-determining resonant circuit caused by mechanical displacements, temperature, pressure, humidity, and time are a completely separate problem which will not be discussed here.

The principles independently arrived at for a basic frequency-stable LC oscillator are simply stated:

1. Use a high-quality tuned circuit (high Q).
2. Use a "series-tuned" circuit, which is equivalent to tapping the vacuum-tube input and output across as low impedances as possible in the tuned circuit, while maintaining oscillation.
3. Use an external system of automatic level control, keeping grid and plate swings of the oscillator tube to a volt or so and grid current at zero.
4. Choose a tube having a low $\Delta C/g_m$ ratio.
5. Operate the vacuum tube in as linear a region of its characteristic as possible.

Brief comments on these points follow:

1. In familiar LC oscillator circuits, such as the Colpitts and Hartley, the tube impedances are low, and they are connected across impedances in the LC circuit many times higher in magnitude. The "loaded" Q of the circuit is then only a fraction of the natural or "unloaded" Q, so that any benefits derived from a high-Q circuit are completely wiped out.
2. If an oscillator circuit, no matter how good it may be in circuit design, is poorly constructed, the frequency stability will be poor. Good mechanical design is as important as good circuit design.
3. The *series-tuned oscillator* is defined as follows: an oscillator in which the input and output circuits of the tube amplifier are connected across very low reactances in the tuned circuit, and in which the feedback path from output to input is through *a series-tuned circuit, the reactances of which are substantially greater than those across which the vacuum-tube circuits are connected. The frequency of oscillation is close to the resonant frequency of the series-tuned circuit.*
4. Many writers have chosen to use a tube of the highest possible g_m, even placing tubes in parallel to increase the value of g_m, *based on the assumption of a specified change in, say, tube input capacitance.* In the actual circuit, the change in grid-cathode capacitance is a property of the tube, and in general it is greater in high-g_m tubes than in other types. If tubes change their input capacitance by a given percentage, then least capacitance change takes place in the tube having the lowest input capacitance. The g_m of the tube determines the minimum magnitude of the coupling reactances, which in turn define the factor by which the resulting frequency change is reduced. The tube having the least $\Delta C/g_m$ ratio will therefore give the best frequency stability against changes in tube capacitances.
5. Linear operation is a necessity to prevent frequency changes produced by generation of harmonics. Several writers[1,2] have shown that the presence of harmonics causes a phase shift of the fundamental component. Since this shift will vary with the amplitude of oscillation and the harmonic content, which can change radically with changes in operating level, it must be kept down by linear operation and by a good automatic level control.
6. From the preceding paragraph, it is evident why limiting by grid current is detrimental to frequency stability: harmonics are generated. The harmonic content changes with level, resulting in phase shifts of the fundamental.

17. Stability Considerations. A great deal of attention has been given to the question of stability of frequency of oscillators, to means for improving frequency stability by use of stabilizing reactances, and to studies of the frequency stability as a function of the circuit parameters, notably by Llewellyn,[1] Groszkowski,[2] Heising,[3]

[1] Llewellyn, F. B., Constant Frequency Oscillators, *Proc. IRE*, **19** (12), 2063, December, 1931.
[2] Groszkowski, Janusz, The Interdependence of Frequency Variation and Harmonic Content and the Problem of Constant-frequency Oscillators, *Proc. IRE*, **21** (7), 958, July, 1933.
[3] Heising, R. A., Stability in High-frequency Oscillators, *Proc. IRE*, **31**, 595, 1943.

Mallett,[1] Gladwin,[2] Lukaszewicz,[3] Beurle,[4] and Roberts.[5] Attention here will be confined to circuit arrangements giving improved frequency stability by linear operation of the tube and an accompanying means of amplitude control, rather than by use of compensating reactances.

Jiri Vackar[6] has treated the oscillatory circuit as a four-pole and given a straightforward analysis. Since this paper is not readily available to many readers, a summary covering the essential features from an engineering point of view was given by Clapp[7] and also by Edson.[8]

Basically, Vackar starts from the parallel resonant impedance R_0 of the tuned circuit and uses the fact that, when one branch is tapped at a point having a reactance X_2 to ground, the impedance across the tap to ground is $R_0X_2^2/(X_1 + X_2)^2$. Since the internal reactances of a vacuum tube are inherently capacitative over very wide ranges of frequency, it follows that, if the tuned circuit is tapped on the capacitative branch, the tube reactances are placed in parallel with reactances of like kind, leading to a number of favorable results which have been appreciated by many writers. The resonant impedance of the tuned circuit is

$$R_0 = QX_0 = \omega^2L^2/R_s = 1/\omega^2C_0^2R_s = L/C_0R_s$$

where C_0 = the net total resonant capacitance of the capacitative branch, and R_s = the effective *series* resistance of the loop. To determine the impedances Z_1' and Z_2' presented to the input and output terminals of the tube, it is necessary to know the value of R_0 and the values of the capacitative reactances determining the stepdown impedance ratio.

Fig. 19. Oscillator and equivalent. To oscillate, $g_mX_1X_2 = R_s$.

Clapp[9] presented an analysis of the input impedance of a tube, together with the associated input and output circuit reactances, showing that this combination acts as a negative resistance of value $-g_mX_1X_2$, in series with the (low) circuit reactances X_1 and X_2. If, across the grid and plate tube terminals, an inductance or an inductance and series capacitance are connected, oscillation will take place if $g_mX_1X_2 \geq R_s$ (see Fig. 19).

It is instructive to start from the parallel-resonant circuit analysis and show that the series circuit analysis can be derived directly from it.

Vackar states that the condition for oscillation is $1/g_m = \sqrt{Z_1'Z_2'}$. Applying this to the basic circuit (column 2, Table 1, page 20), we have

$$\begin{aligned}
Z_1' &= R_0(C_vC_1)^2/(C_v + C_1)^2 = R_0C_v^2/C_1^2(1 + C_v/C_1)^2 \cong R_0(C_v/C_1)^2 \\
Z_2' &= R_0(C_vC_2)^2/(C_v + C_2)^2 = R_0C_v^2/C_2^2(1 + C_v/C_2)^2 \cong R_0(C_v/C_2)^2 \\
g_m &= 1/\sqrt{Z_1'Z_2'} = (C_1/R_0)(C_1/C_v)(C_2/C_v) \\
&= R_s\omega^2C_v^2(C_1/C_v)(C_2/C_v) \\
&= R_s\omega^2C_1C_2 \\
&= R_s/X_1X_2
\end{aligned}$$

as given by Clapp, neglecting small correction terms.

[1] Mallett, E., Frequency Stabilization of Valve Oscillators, *J. IEE*, **68**, 578, 1930.

[2] Gladwin, A. S., Stability of Oscillation in Valve Generators, *Wireless Eng.*, August, 1955, p. 206; September, 1955, p. 246; October, 1955, p. 272; November, 1955, p. 297; Constant Frequency Oscillators, *Wireless Eng.*, January, 1956, p. 13.

[3] Lukaszewicz, L., Frequency Stabilization of Valve Oscillators in Respect of Grid Current, Amplification Factor and Load, *Bull. de l'academie polonaise de sciences*, class IV, II (4), 181, 1954.

[4] Beurle, R. L., A Comparison of the Noise, Random Frequency and Amplitude Fluctuations in Different Types of Oscillators, *Proc. IEE* (British), **103**, Part B, 182, March, 1956.

[5] Roberts, W. V. B., The Limits of Inherent Frequency Stability, *RCA Rev.*, **4** (4), 478, April, 1940.

[6] Vackar, Jiri, LC Oscillators and Their Frequency Stability, *Tesla Tech. Repts.*, Czechoslovakia, December, 1949, pp. 1-9.

[7] Clapp, J. K., Frequency Stable LC Oscillators, *Proc. IRE*, **42** (8), 1295, August, 1954.

[8] Edson, W. A., "Vacuum Tube Oscillators," pp. 170-172, Wiley, 1953.

[9] Clapp, J. K., An Inductance-Capacitance Oscillator of Unusual Frequency Stability, *Proc. IRE*, **36** (3), 356, March, 1948.

Vackar's result is of engineering interest because it emphasizes the fact that the tube is connected across reactances which, as stepped down by the tapping, are *very low* compared with the resonant impedance. However, in stating his results in terms of Q, interpretation is clouded since assumption that Q varies in a certain manner or remains constant as the frequency is varied may be quite unjustified. In the series circuit analysis, the parallel-resonant impedance is of no direct interest, and Q comes into the picture only incidentally as R_s changes. There is no interest in the impedance stepdown ratio as such, since only the magnitudes of the reactances, across which the tube is connected, are involved.

The series analysis is based on the vacuum tube with the associated input and output reactances, as shown in Fig. 19.

The effect of a change ΔC_1 in the input capacitance of the tube, connected across the tapped impedance Z_1' of the tuned circuit, causes a detuning equivalent to a change ΔC_0 in the net tuned-circuit capacitance C_0 such that

$$\Delta C_0/\Delta C_1 = Z_1'/R_0 \tag{1}$$

where R_0 = parallel impedance of the tuned circuit.

If the tuned-circuit capacitance is C_0, a change ΔC_0 in this capacitance causes the fractional frequency change

$$\begin{aligned}\Delta f/f_0 &= \Delta C_0/2C_0 \\ &= (Z_1'/2R_0)(\Delta C_1/C_0)\end{aligned} \tag{2}$$

Similar considerations apply for changes in the output capacitance of the tube. For *equal* changes in *either* input or output tube capacitances, the minimum frequency change occurs when $Z_1' = Z_2' = 1/g_m$, and is

$$\begin{aligned}\Delta f/f_0 &= (1/2R_0C_0)(\Delta C_1/g_m) \\ &= (\omega_0/2Q)(\Delta C_1/g_m) \\ &= (2\pi f_0/2Q)(\Delta C_1/g_m)\end{aligned} \tag{3}$$

Frequently, the change in tube output capacitance is less than that of the input capacitance. In such cases, slightly improved stability can be obtained by making $Z_1' \neq Z_2'$. Then

$$\begin{aligned}\Delta f/f_0 &= (1/2R_0C_0)(\Delta C_1/g_m)(2/\sqrt{k}) \\ &= (\omega_0/2Q)(\Delta C_1/g_m)(2/\sqrt{k}) \\ &= (2\pi f_0/2Q)(\Delta C_1/g_m)(2/\sqrt{k})\end{aligned} \tag{4}$$

where $k = \Delta C_1/\Delta C_2$.

Equation (4) is very useful since it gives the frequency-stability coefficient immediately when the quality of the tuned circuit and the $\Delta C_1/g_m$ ratio of the tube are known.

If ΔC_1 were independent of g_m, the tube having the highest g_m would give the best frequency stability; this conclusion has been reached by several writers. In practice, however, the tubes having the larger values of g_m have also the larger values of C_1 and the larger values of ΔC_1. The choice of a tube having very low capacitance and small changes of tube capacitances associated with only moderate values of g_m will frequently result in substantially lower $\Delta C_1/g_m$ ratio with consequent improved frequency stability. This is particularly true of the secondary changes in tube capacitances such as those caused by changes in temperature of tube structure and by changes in emission of the cathodes.

Equation (4) also indicates that the frequency stability is independent of the L/C_0 ratio of the tuned circuit. The stability depends only on the quality of the tuned circuit (first factor) and, for a given value of ΔC_1, on the g_m of the tube (second factor). The latter term determines the minimum possible coupling between the driving circuit and the controlling circuit.

In Table 1, several frequency-stable LC oscillators are shown, in progressive order of adaptation of the coupling means to the tube, to give wider frequency ranges and more nearly constant amplitude over the tuning band. The equivalent circuits are shown, where a single variable capacitance is used but effectively acts as if it were a two- or three-gang unit.

While the basic oscillation circuit is shown, details of connection to the tube are omitted. Depending upon the position of the rotor of the variable capacitor, various

ground points may be chosen so as to ground the rotor. Frequently, a grounded-plate oscillator is desirable, since no d-c voltage then appears on the tuned-circuit elements.

The positions of the inductor and variable capacitor can be interchanged without difficulty if advantage in grounding the capacitor results.

Two examples of complete oscillators are shown in the bottom of the chart; one (column 2) is a grounded-plate version and one (column 5) is a grounded-cathode circuit. In the first, the d-c voltage is removed from the tuned circuit while using series-connected plate supply. In the second the d-c voltage is removed from the tuned circuit while using parallel-connected plate supply.

In each column, the g_m required for oscillation is given in terms of C_v, the tuning capacitance. If the required g_m decreases as C_v decreases, the amplitude of oscillation will increase as the frequency is increased. In other words, if the circuit is tuned toward lower frequencies, the amplitude decreases, and if C_v becomes large enough, oscillation ceases. Such circuits are shown in columns 1, 4, and 6. The circuit of column 2 acts in the reverse sense, with oscillation ceasing if C_v becomes sufficiently small. The circuits of columns 3 and 5 show more nearly constant amplitude with tuning.

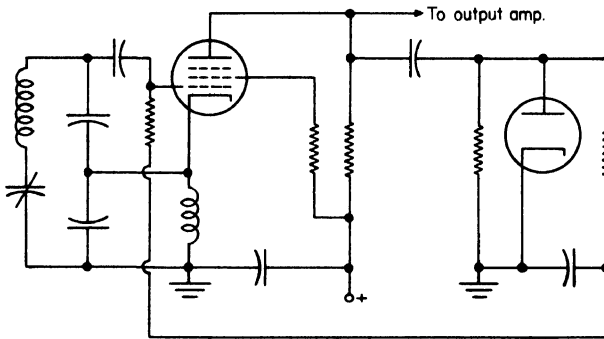
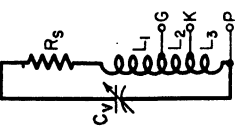
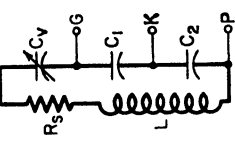
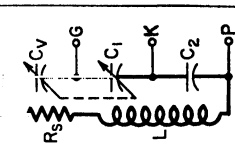
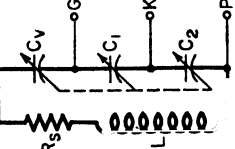
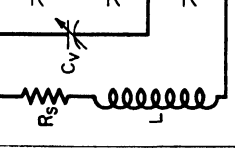
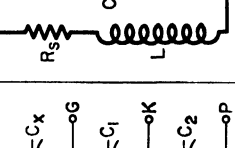
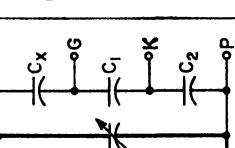



FIG. 20. Clapp-Gouriet oscillator, a modified Colpitts.

All the above statements are strictly true only when R_s remains constant over the tuning range. Since R_s can *decrease* as frequency *increases* in the low-frequency ranges (say 10 to 1,000 kc) and can *increase* in the higher-frequency ranges (say 500 kc to a few megacycles), the performance of any of the oscillator circuits can be modified appreciably. In the range from a few megacycles to several tens of megacycles, R_s generally decreases gradually as the frequency increases as long as the natural frequency of the inductance is not approached too closely.[1] The circuit of column 7 has provision for varying the coupling by a combination of variable and fixed capacitances C_2 and C_3. If C_3 is large compared with C_2, the performance is similar to that of the circuits of columns 4 and 6. If C_3 is of the same order as C_2, the performance approaches that of the circuits of columns 3 and 5. By proportioning of these two capacitances, adjustment can be made to obtain nearly constant amplitude in spite of variation of R_s over the tuning range.

In each column, the originator is named, with references to publication. The usually effective tuning range is given at the bottom of the column.

For operation at a single frequency, or over a very limited frequency range (not greater than 1.03:1, for example), all the circuits shown, if properly adjusted, would give identical stabilities. Consequently, the simplest circuit (column 2) would be the one to use, since no advantage results from using more elements.

As the tuning range is extended, the more complicated circuits give much more nearly constant amplitude over the tuning range and the added elements are well

[1] Field, R. F., Reduction of Losses in Air-cored Coils, *Proc. Symposium on Improved Quality Electronics Components*, Washington, D.C., May, 1950.

Table 1. Frequency-stable *LC* Oscillators

1	2	3	4	5	6	7
Lampkin[a]	Gouriet[b]-Clapp[c]			Vackar circuit A[d]	Seiler[e]	Vackar[d] circuit B
One variable element	One variable element	Two variable elements	Three variable elements	One variable element equivalent of two	One variable element equivalent of three	Combines variable C_2 of Col. 4 with fixed C_2 of Col. 3
$L_2 \ll L_1$, $L_3 \ll L_1$ $g_m = \frac{R_s}{\omega^2 L_2 L_3}$ $\omega^2 = \frac{1}{LC_v}$ $g_m = \frac{R_s L_1}{L_2 L_3} C_v$ g_m req decreases as f increases. f range 1.8:1	$C_v \ll C_1$, $C_v \ll C_2$ $g_m = R_s\omega^2 C_1 C_2$ $\omega^2 \doteq \frac{1}{LC_v}$ $g_m = \frac{R_s C_1 C_2}{LC_v}$ g_m req increases as f increases. f range 1.2:1	$C_v \ll C_1$, $C_v \ll C_2$ $C_1 = k_1 C_v$ $g_m = R_s\omega^2 C_1 C_2$ $\omega^2 = \frac{1}{LC_v}$ $g_m = \frac{R_s C_1 C_2}{LC_v}$ $= \frac{R_s k_1 C_v C_2}{LC_v}$ $= \frac{R_s k_1 C_2}{L}$ g_m req = const as f increases (if R_s = const). f range = 1.8:1	$C_v \ll C_1$, $C_v \ll C_2$ $C_1 = k_1 C_v$, $C_2 = k_2 C_v$ $g_m = R_s\omega^2 C_1 C_2$ $\omega^2 \doteq \frac{1}{LC_v}$ $g_m = \frac{R_s C_1 C_2}{LC_v}$ $= \frac{R_s}{L} k_1 k_2 C_v$ g_m req decreases as f increases. f range 1.8:1	Equivalent of above. Compare with Col. 3	Equivalent of above. Compare with Col. 4	Equivalent of above. Compare with Cols. 3 and 4.

Oscillator (circuit Col. 2)	Oscillator (circuit Col. 5)			
Q = series-resonant crystal		$C_A = C_a + C_x = C_x + C_v \frac{C_x}{C_T}$ $C_B = C_b + C_1 = C_1 + C_v \frac{C_1}{C_T}$ $C_T = \frac{C_x C_1}{C_x + C_1}$ $\frac{C_a C_b}{C_a + C_b} = C_v$ $C_v \ll C_1, C_v \ll C_2$ $g_m = \frac{R_s \omega^2 C_B C_2}{LC_A}$ $C_B = k_1 C_A$ $g_m = \frac{R_s k_1 C_A C_2}{LC_A}$ $= \frac{R_s}{L} k_1 C_2$ g_m req = const as f increases (if R_s = const). f range 1.8:1	$C_A = C_x + C_a = C_x + C_v \frac{C_x}{C_T}$ $C_B = C_1 + C_b = C_1 + C_v \frac{C_1}{C_T}$ $C_C = C_2 + C_e = C_2 + C_v \frac{C_2}{C_T}$ $C_T = \frac{C_x C_1 C_2}{C_x C_1 + C_x C_2 + C_1 C_2}$ $\frac{C_a C_b C_e}{C_a C_b + C_a C_e + C_b C_e} = C_v$ $C_a \ll C_b, C_a \ll C_e$ $C_B = k_1 C_A, C_C = k_2 C_A$ $g_m = R_s \omega^2 C_B C_C$ $\omega^2 \doteq \frac{1}{LC_A}$ $g_m = \frac{R_s k_1 C_A k_2 C_A}{LC_A}$ $= \frac{R_s}{L} k_1 k_2 C_A$ g_m req decreases as f increases. f range 1.8:1	Note that C_2, C_3 replace C_2 of Cols. 3 and 4 $C_A = C_x + C_a = C_x + C_v \frac{C_x}{C_T}$ $C_B = C_1 + C_b = C_1 + C_v \frac{C_1}{C_T}$ $C_C = C_2 + C_e = C_2 + C_v \frac{C_2}{C_T}$ $C_T = \frac{C_x C_1 C_2}{C_x C_1 + C_x C_2 + C_1 C_2}$ $\frac{C_a C_b C_e}{C_a C_b + C_a C_e + C_b C_e} = C_v$ $C_a \ll C_b, C_a \ll C_e, C_2 \ll C_3$ $C_B = k_1 C_A, C_C = k_2 C_A$ $g_m = R_s \omega^2 C_B \left(\frac{C_e C_3}{C_e + C_3}\right)$ $\omega^2 \doteq \frac{1}{LC_A}$ $g_m = \frac{R_s}{L} k_1 \frac{k_2}{1 + \frac{k_2 C_A}{C_3}} C_A$ If C_3 very large, g_m req same as Col. 4. If C_3 smaller, g_m req can be adjusted toward value given in Col. 3. f range 2.5:1

[a] Lampkin, G. F., An Improvement in Constant-frequency Oscillators, *Proc. IRE*, **27** (3), 199, March, 1939.
[b] Sandeman, E. K., "Radio Engineering," Chapman & Hall, 1947. Gouriet, G. G., High-stability Oscillator, *Wireless Engr.*, April, 1950, pp. 105–112.
[c] Clapp, J. K., An Inductance-Capacitance Oscillator of Unusual Frequency Stability, *Proc. IRE*, **36** (3), 356, March, 1948. Roberts, W. A., Discussion, Correspondence Section, *Proc. IRE*, **36** (10), 1261, October, 1948.
[d] Vackar, Jiri, LC Oscillators and Their Frequency Stability, *Tesla Tech. Repts.*, Czechoslovakia, December, 1949, pp. 1–9.
[e] Seiler, E. O., A Low-C Electron-coupled Oscillator, *Q.S.T.*, November, 1941, pp. 26–27.

justified. The circuit of column 5 is first choice because a single variable element is used, which rules out the circuit of column 3.

The added complexity of the circuit of column 7 is seldom justified. It would be used only where a very wide tuning range was a prime requirement.

18. Amplitude Control. Gouriet[1] shows a simple amplitude regulator applied to the oscillator of column 1 (but a similar arrangement can be used with any of the other circuits). A resistance is added in series with the plate to provide an r-f voltage which is applied to a diode rectifier. A germanium diode could equally well be used. The rectified voltage is filtered and applied at the base of the grid resistor R_g.

A more complete circuit is shown in Fig. 21. Here, an amplifier is added to provide greater drive of the diode. With the amplifier, the circuit can maintain practically a constant and low level of about 0.1 to 0.01 volt. An output amplifier is used in

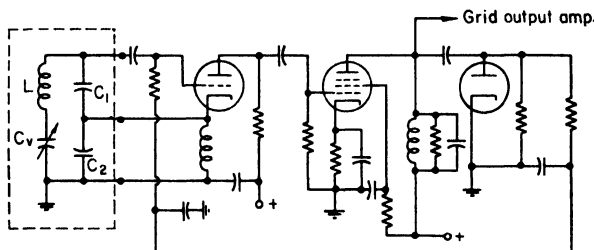



FIG. 21. Practical v-f-o oscillator for transmitter. LC_v is in a temperature-controlled space. For LC_v a crystal may be substituted with appropriate values of C_1 and C_2. Plug-in coils may be used.

practical applications. Such oscillator circuits have been used in master oscillators of transmitters. The performance attained is indicated by the following figures:

1. Plate-voltage change = ±20 per cent; Δf resulting = $\pm 2.5/10^6$
2. Filament-voltage change = ±5 per cent; Δf resulting = $\pm 2.5/10^6$
3. Stability with time

1 hr.	$\Delta f = \pm 1/10^6$
1 day	$\Delta f = \pm 10/10^6$
1 month	Δf = with reset accuracy of $\pm 30/10^6$

A sensitive means of testing or comparing LC frequency-stable oscillators to determine how free the frequency of oscillations is from any variation due to tube voltages is afforded by the locking system described in Sec. 61. Where an ordinary Hartley or Colpitts oscillator will lock over a range from several tens to several hundred parts per million, a well-adjusted frequency-stable oscillator will show no locking whatever.

19. General-purpose Oscillators. Oscillators for general application as power sources in the a-f and r-f ranges can be constructed with LC tuned circuits as the principal frequency-determining elements. The circuits shown in Fig. 22 are shown as triode circuits, but they may be used as multigrid circuits if the appropriate elements are added. Also by the use of multigrid tubes and proper circuit manipulation, each of these circuits can be changed to provide an electron-coupled output circuit. An electron-coupled Colpitts oscillator is shown at f in Fig. 22.

20. Choice of Oscillator Circuit. The actual circuit used must depend on the application for which it is intended. Several representative uses of the several oscillator circuits as shown in Fig. 22 may be of interest in helping define the field of application of the various circuits.

The tuned-grid oscillator is frequently used in radio receivers as a regenerative detector or as the local oscillator in superheterodyne receivers. It can be made to

[1] Gouriet, G. G., High-stability Oscillator, *Wireless Engr.*, April, 1950, pp. 105–112.

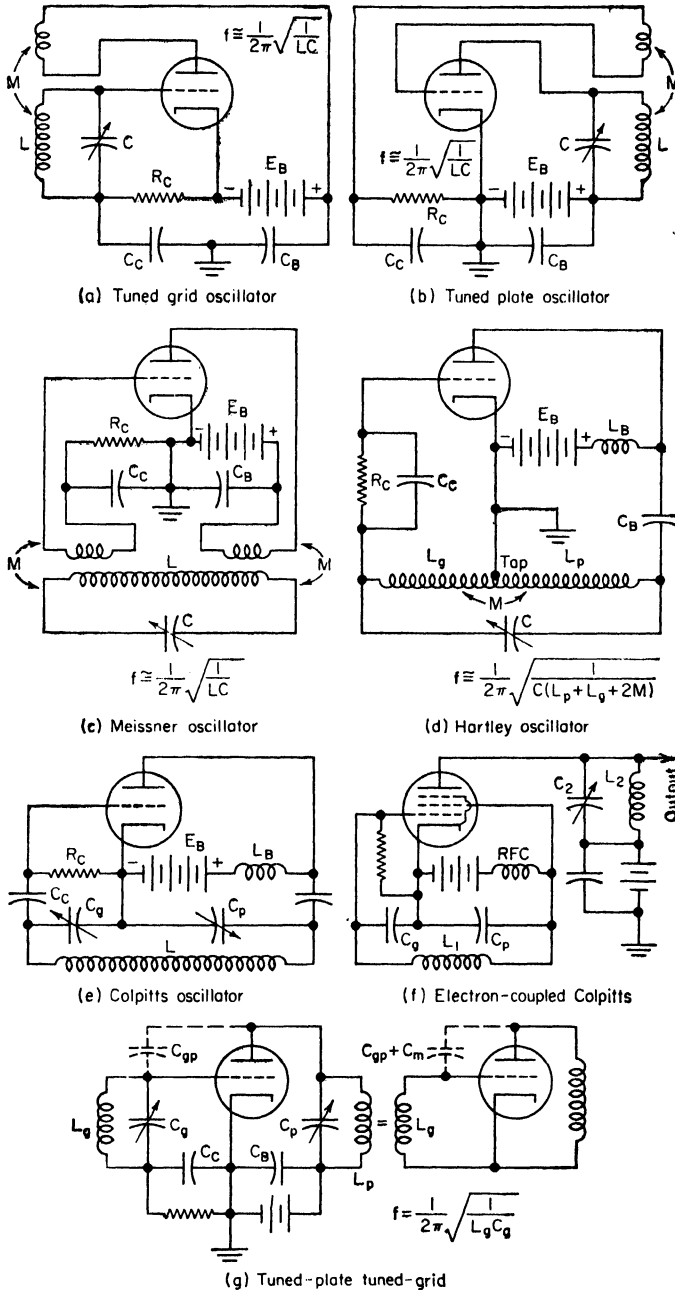



FIG. 22. Types of feedback oscillators.

work over a wide frequency range by changing or switching coils, and it requires only one capacitor for tuning.

The tuned-plate oscillator is often used in test equipment where wide frequency ranges are to be covered with relatively low power output.

The Meissner oscillator has been used in various applications at both high and low power. It is almost impossible to make this oscillator perform properly over anything more than a very narrow frequency range, since the many possible parasitic resonances must be suppressed if successful operation is to result. Nevertheless, this circuit has been used at very high power level, in a cyclotron, because the reduction in fed-back driving power to the grid circuit is automatic when the dee resonator arcs over, thus preventing catastrophic damage to the dees when an arc occurs. Another application of the Meissner circuit substituted a high-Q transmission-type cavity resonator for the usual lumped-circuit elements. With a good microwave triode (W. E. 416-B), a stable oscillator was constructed, using this circuit, for use with a cesium-beam apparatus. The oscillator operated near 3,064 Mc and a diode-type harmonic generator was used to obtain output near 9,192 Mc.

The Hartley oscillator is especially useful at low frequencies from the audio range up to approximately 40 to 50 Mc. At low frequencies, it is possible to maintain stable feedback conditions because of the tapped inductance even though circuit Q and impedance may vary widely with tuning. This circuit is used, for example, in several widely used grid-dip oscillators where freedom from variation in feedback is essential, and also in a general-purpose test oscillator covering the range from 5 kc to 50 Mc. It is also used in many transmitters in the low- and medium-frequency bands, where crystal frequency control is not essential, usually followed by an amplifier.

The Colpitts oscillator is widely used because of its relatively good stability and low-harmonic output. It is used almost exclusively at frequencies above 100 Mc where almost any tunable oscillator tends to become a Colpitts circuit with the tube capacitances as the limiting values of C_g and C_p.

The tuned-plate tuned-grid oscillator is sometimes used to provide different impedance levels in the plate and grid circuits where, for example, high-voltage tuning capacitors may be available for only part of the circuit. It is then possible to use the *TP-TG* circuit with a large low-voltage-rating capacitor to tune the grid circuit of an oscillator to supply reasonable amounts of power with fairly good stability as, for example, in dielectric heating applications. Push-pull oscillators using resonant parallel lines as tuning elements for plate and grid circuits with cathode grounded are examples of the tuned-plate tuned-grid oscillator circuit.

The grounded-grid oscillator is similar in operation to the *TP-TG* oscillator. It is used principally in the u-h-f band since (1) disk-seal tubes are available which are well suited for use in this circuit, and (2) almost no other circuit is as satisfactory in this range. Coaxial-line resonators are usually employed with disk-seal tubes, wide tuning ranges being afforded by the adjustment of the coaxial line length with sliding plungers.

HIGH-FREQUENCY OSCILLATORS

At frequencies above about 100 Mc, the performance of conventional oscillators is significantly affected by the following factors:

1. Transit time
2. Lead and electrode inductance and capacitance
3. Circuit losses

The maximum frequency at which oscillations can be obtained with a particular tube and circuit is usually limited by one or more of the above factors.

21. Transit Time. At high frequencies, the time required for an electron to travel between various tube electrodes, which is referred to as transit time, cannot be

assumed to be negligible. The velocity v of an electron is proportional to the square root of the potential V through which it has traveled.

$$v = 5.93 \times 10^7 \sqrt{V} \quad \text{cm per sec}$$

Since electrons start at approximately zero velocity at the cathode and since the cathode-grid voltage is usually small, the transit time between the cathode and grid in conventional tubes is relatively large. The transit time between grid and plate, however, is usually much smaller than the cathode-grid transit time even though the distance is greater because electrons enter the space with an appreciable velocity and because the plate potential is much higher than that of the grid.

Two effects of transit time are of prime importance: (1) the reduction of and phase shift in the effective transadmittance (at low frequencies equal to the transconductance) of the tube and (2) the increase in grid loading. The variation in the magnitude and phase of the transadmittance as a function of transit angle[1] is shown in Fig. 23. The transit angle is the ratio of the electron-transit time to the period of the signal expressed as an angle. Since the cathode-grid transit time is usually the largest, it will have a significant effect at a much lower frequency than will the grid-plate transit time. Note that the effective transadmittance initially decreases from the d-c value as the frequency is increased and that its phase angle becomes increasingly lagging.

FIG. 23. Magnitude and phase of the effective transadmittance as a function of transit angles. (*From Llewellyn and Peterson.*)

At low frequencies, the current induced in the grid circuit by the electron stream leads the grid voltage by 90 deg and hence effectively increases the grid-to-cathode capacitance. At high frequencies, transit time causes the phase angle of the induced current with respect to the grid voltage to be less than 90 deg and hence results in a conductive component in the input admittance.[2] The input conductance increases approximately as f^2, and therefore the parallel input resistance decreases as f^2, as shown[3] in Fig. 24. As a result, the grid losses increase at the higher frequencies and

[1] Llewellyn, F. B., and L. C. Peterson, *Proc. IRE*, **32**, 144, March, 1944.
[2] Spangenberg, K. R., "Vacuum Tubes," Sec. 16.5, McGraw-Hill, 1948. Ferris, W. R., *Proc. IRE*, **24**, 82, January, 1936. North, D. O., *Proc. IRE*, **24**, 108, January, 1936.
[3] Lavoo, N. T., *Proc. IRE*, **35**, 1248, November, 1947.

can be a serious limitation on the maximum frequency at which oscillations can be obtained.

22. Lead-reactance Effects. Lead inductances and electrode capacitances can also be important factors. Inductance in the cathode lead[1] primarily causes the grid input resistance to decrease as f^2 in the same manner as does the electron-transit time.

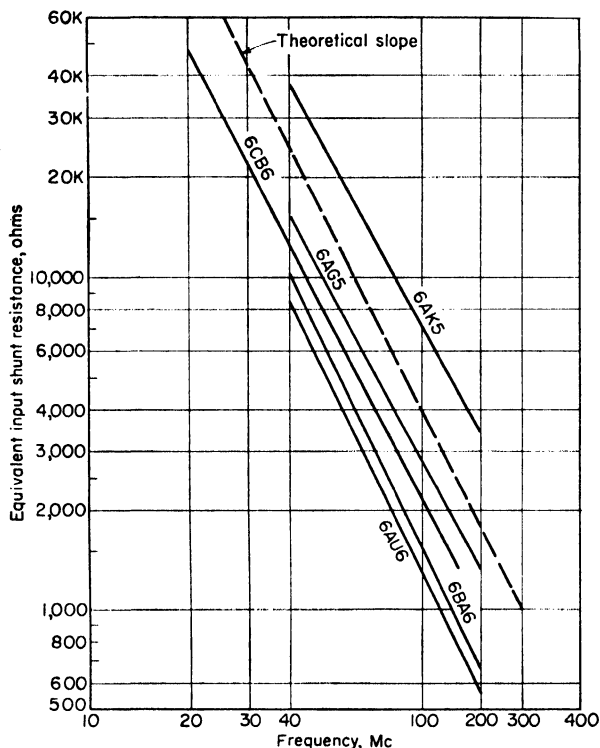



FIG. 24. Input resistance of typical receiving tubes under normal operating conditions as a function of frequency.

The combination of internal lead inductance and electrode capacitance limits the maximum frequency at which oscillations can be obtained with lumped circuits. This limit is usually reached when the inductive reactance of the leads equals the reactance of the interelectrode capacitances effectively in series and in parallel, respectively, with the resonant circuit as shown in Fig. 25. Under these conditions, a parallel-resonant circuit will appear across the actual tube electrodes with the tube external circuit connections short-circuited. Somewhat higher oscillating frequencies can be obtained with lumped circuits by the connection of a capacitor across the tube terminals in place of the resonant or short circuit.[2]

[1] "Input Admittance of Receiving Tubes," RCA Application Notes, No. 188, Radio Corporation of America, April, 1947. Strutt, M. J. O., and A. van der Ziel, *Proc. IRE*, **26**, 1011, August, 1938. Sarbacher, R. I., and W. I. Edson, "Hyper and Ultra-high Frequency Engineering," pp. 431–436, Wiley, 1943.

[2] Pettit, J. M., *Proc. IRE*, **38**, 633, June, 1950.

Losses due to skin effect in the conductors, whose resistance increases as the square root of frequency, and dielectric losses in the glass envelope also can be serious at high frequencies.

23. High-frequency Triode Oscillators. Many conventional receiving-type tubes, such as the 6AK4, 6AK5, 6C4, 6J4, 6J6, and 9002, can be used in low-power tunable oscillators with lumped *LC* resonant circuits at frequencies up to about 500 Mc. Other conventional-type tubes designed for h-f use, such as the 6F4, 6AF4, and 6AJ4, will oscillate at frequencies up to about 1,000 Mc with special types of lumped resonant circuits. Special types of triode tubes, such as the 2C40, 2C39, 2C37, 5675, and 416A, are suitable for use as oscillators at frequencies up to about 4,000 Mc.

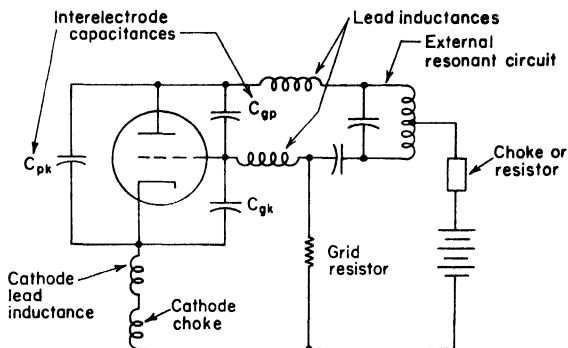
FIG. 25. Typical low-power, high-frequency triode oscillator showing interelectrode capacitances and lead inductances.

High-power tubes of conventional construction have top frequency limits in the 50- to 100-Mc frequency range while special types of high-power triodes are available for use up to about 1,000 Mc.

In most low-power oscillators for use below 1,000 Mc, the modified Colpitts circuit of Fig. 25 is used in which the feedback required for oscillation is obtained by means of the interelectrode tube capacitances.[1] In this arrangement, the cathode and heater are "floating" with respect to ground to eliminate the effect of cathode-lead inductance and external capacitances on the operation. Although it is not possible to make the impedance between cathode and ground infinite as is required for a perfect "floating" cathode, satisfactory performance over a wide band of frequencies can be obtained by the use of properly designed chokes in the heater and cathode leads.[2]

The inductance in variable capacitors used in the resonant circuit must be held at a minimum to obtain optimum high-frequency performance. Balanced, split-stator variable capacitors perform well at high frequencies since low-inductance connections can be made to the stator segments and the connection to the rotor, which usually has a relatively large inductance, is eliminated. With the balanced capacitor, the inductor usually consists of one or more straps or bars connecting the two stator sections.

LC resonant circuits with tuning ranges of the order of 5 to 1 can be obtained if the rotor of the variable capacitor is used to reduce the parallel inductance in the circuit by eddy-current shielding as the capacitance is decreased. One type of circuit which performs this function is the *butterfly* circuit.[3] A symmetrical butterfly and its equivalent circuit are shown in Fig. 26. As the rotor leaves the stator, it reduces the area through which the magnetic field set up around the inductance arm can pass and thus decreases the inductance. Figure 27 is a symmetrical butterfly oscillator which operates over the frequency range from 250 to 920 Mc.

24. Oscillators for 1,000 Mc and Above. At frequencies above 1,000 Mc, lumped-circuit resonators become impractical since they must be made very small and usually have low *Q*, and the magnitude and distribution in the circuit of the stray inductances and capacitances have a drastic effect on the performance. At these frequencies,

[1] Kamphoefner, F. J., *Proc. IRE*, **38**, 630-632, June, 1950.
[2] Radio Research Laboratory Staff, "Very High Frequency Techniques," Sec. 30-9, McGraw-Hill, 1947.
[3] Karplus, E., *Proc. IRE*, **33**, 426, July, 1945. Radio Research Laboratory Staff, *op. cit.*, Secs. 28-7 to 28-16; Chap. 30.

distributed-parameter resonators, such as coaxial-line or cavity resonators, are useful since their resonant frequencies can be accurately calculated, they have high Qs, and they are reasonable in size. A length of coaxial line with movable short-circuiting plunger is a commonly used type of resonator.

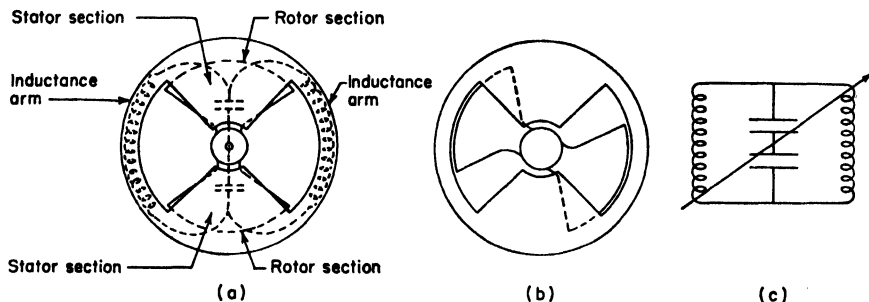



FIG. 26. (*a*) Symmetrical butterfly resonator with rotor set for minimum resonant frequency. No eddy-current shielding of the inductance arms by the rotor is obtained with this setting. (*b*) Butterfly set near middle of tuning range. Appreciable reduction in inductance due to eddy-current shielding is secured. (*c*) Equivalent circuit of butterfly resonator.

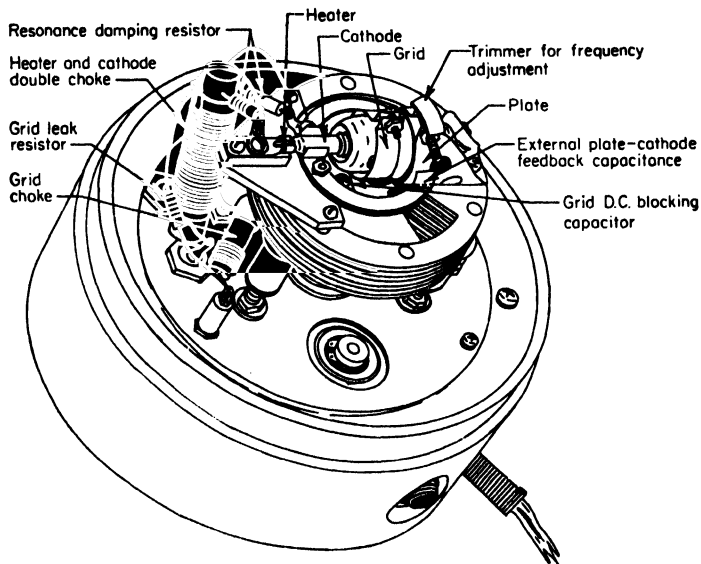



FIG. 27. Butterfly oscillator for 250 to 920 Mc.

To realize the advantages of the coaxial-line resonators, the lead inductance in the connections from the oscillator tube to the coaxial-line circuit must be very small. To achieve this, high-frequency tubes are constructed with the leads made in the form of disks or cylinders which pass directly through the glass and which actually become part of the coaxial-line circuit. In the 5675 pencil triode,[1] the cathode, grid, and plate are cylindrical structures with the grid connection made by means of a

[1] Rose, G. M., D. W. Power, and W. A. Harris, *RCA Rev.*, **10**, 321, September, 1949.

flange and the cathode and plate connections made by means of cylindrical terminals, as shown in Fig. 28. The electrode spacing and hence the transit time are small and operation up to about 3,000 Mc with power outputs of about ½ watt at 1,700 Mc and 50 mw at 3,000 Mc is possible.

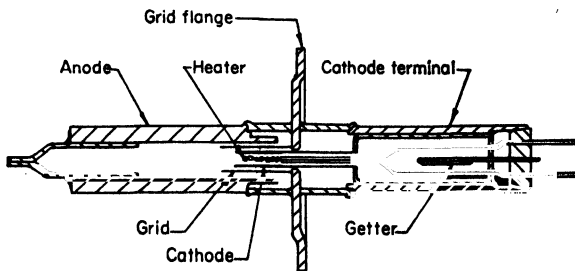
Fig. 28. Cross section of pencil-type tube.

Still higher-frequency tubes are constructed using the *disk-seal, planar-grid*, or *lighthouse* principle in which the cathode, grid, and plate are all parallel-plane electrodes as shown in Fig. 29. Types 2C40 and 2C43 lighthouse tubes[1] and the 2C37 *rocket* tube make satisfactory low-power oscillators at frequencies up to about 3,500 Mc. Power outputs of the order of 40 watts at 1,000 Mc and 18 watts at 2,500 Mc can be obtained with inverted lighthouse tubes, such as the type 2C39. In this type of tube, the plate is at the large end of the tube and hence can be readily cooled.

Fig. 29. Cross section of type 2C40 lighthouse tube.

Still higher-power disk-seal tubes are available for use up to about 1,200 Mc. Tubes such as the 5588 and 4CX250K are capable of generating of the order of 100 watts of c-w power at 1,000 Mc with about 30 per cent efficiency. Other tubes, which are usually used as power amplifiers, such as the 6448, produce c-w outputs of up to 11 kw at frequencies up to 900 Mc.

The extreme in close-spaced, microwave triode tubes[2] is the 416-A shown in Fig. 30. This tube has a cathode-grid spacing of 0.0006 in., a grid-wire diameter of 0.0003 in., a grid-plate spacing of 0.012 in., and a transconductance of 50,000 μmhos. The tube will function as an oscillator considerably above 4,000 Mc.

25. Line Oscillators. Coaxial-line oscillators can be constructed which use only a single resonator[3] in a circuit similar to that shown for lumped-circuit oscillators in Fig. 25. In this arrangement, the tube is connected across one end of the coaxial line, with the grid connected to the outer conductor and the plate to the inner, and a short circuit is placed at the other end of the line. With most high-frequency tubes, the plate-cathode capacitance is usually too small to produce the amount of coupling

[1] McArthur, E. D., *Electronics*, **18**, 98, February, 1945. Reich, H. J., P. F. Ordung, H. C. Krauss, and J. C. Skalnik, "Microwave Theory and Techniques," Chap. 11, Van Nostrand, 1953.

[2] Morton, J. A., and R. M. Ryder, *Bell System Tech. J.*, **29**, 496, October, 1950. Bowen, A. E., and W. W. Mumford, *Bell System Tech. J.*, **29**, 531, October, 1950.

[3] Radio Research Laboratory Staff, *op. cit.*, Chap. 32.

required for oscillation, and additional external coupling must be provided. If a wide tuning range is desired, it may be difficult to design a coupling element which performs satisfactorily over the whole frequency range. Also the top frequency is usually limited to the natural resonant frequency of the tube, which is that obtained with the short circuit as close as possible to the tube with the coaxial line operating in its quarter-wave mode. Usually the frequency limit is around 1,800 Mc.

FIG. 30. Bell System type 416-A "closed-spaced" triode.

For operation at higher frequencies and for more efficient operation at lower frequencies, two resonant lines are used. One is connected between grid and plate and the other between grid and cathode[1] as shown in Fig. 31. The grid-cathode line can be folded back on the grid-plate line to save space, as shown in Fig. 32. With two resonators, the proper feedback conditions can be produced to force the circuit to oscillate with the cavities in the $\frac{3}{4}\lambda$ mode rather than the $\lambda/4$ mode. Under these conditions, the frequency range is not limited by the natural resonant frequency of the tube and much-higher-frequency operation can be obtained. As in the single-cavity oscillator, the plate-cathode capacitance is usually too small to provide the proper coupling; therefore a small amount of additional coupling must be provided between the cavities. Small loops or probes are satisfactory for this purpose.

With movable short circuits, wide frequency ranges can be obtained with coaxial-line oscillators using disk-seal tubes. A range from 300 to 3,000 Mc is realizable with the proper circuit design.[2] The short-circuiting elements in the coaxial line can be of the finger type, in which actual physical contact is made between the plunger and the line, or they can be of the noncontacting or choke type,[3] in which no actual connections are made but a very low r-f impedance is produced at the face of the plunger. Figure 33 shows several types of noncontacting plungers along with their performance characteristics. Note the excellent broad-band characteristics of the S-type plunger.

FIG. 31. Equivalent circuit of two-resonator coaxial-line oscillator.

[1] Stephenson, J. G., *IRE Convention Record*, Part 8, p. 106, 1954.
[2] Radio Research Laboratory Staff, *op. cit.*, Chaps. 14, 15.
[3] Huggins, W. H., *Proc. IRE*, **35**, Part 1, p. 906; Part 2, p. 1085; Part 3, p. 1324, September, October, and November, 1947.

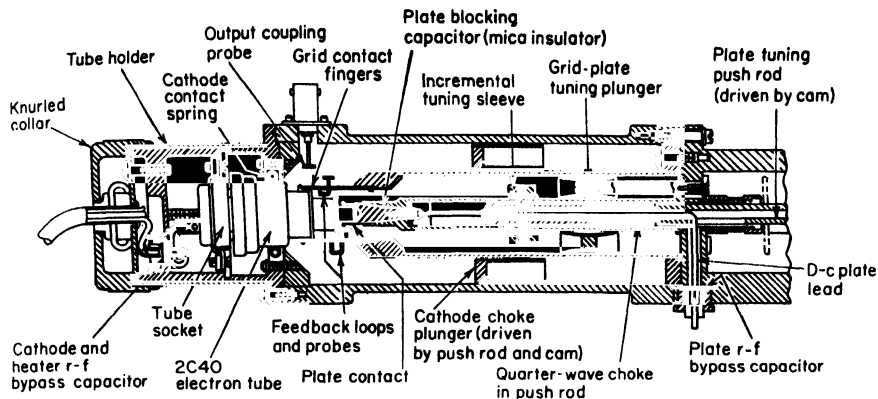



FIG. 32. Coaxial-line oscillator with reentrant resonators. (*From Stephenson.*)

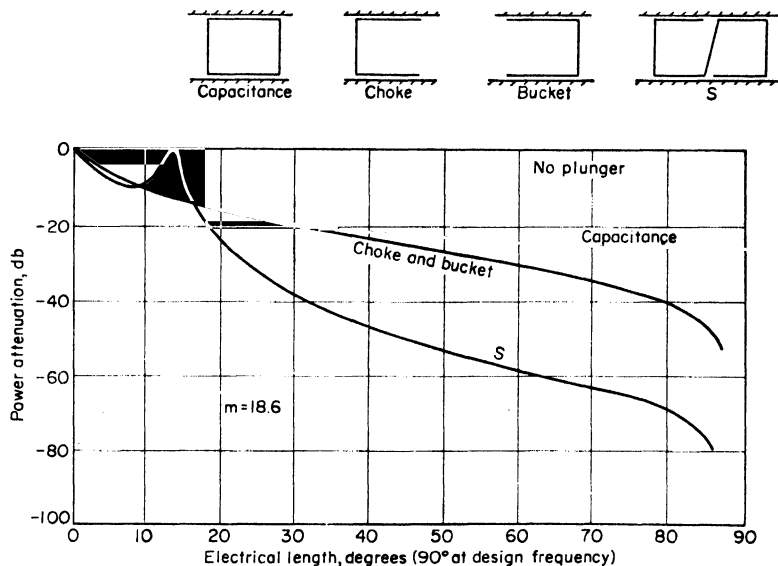



FIG. 33. Rear-cavity power loss of basic noncontacting plungers. The ratio of characteristic impedance of main coaxial line to that of the gaps on either side of the plungers m is assumed to be 18.6 in the graph.

High-power triode oscillators can be obtained by the use of a number of disk-seal tubes mounted in an annular resonator.[1]

The *Resnatron*[2] is a very-high-power tetrode oscillator in which the tube electrodes are actually an integral part of the coaxial-line resonators, as shown in Fig. 34. Power outputs of the order of 60 kw c-w have been obtained over the 340- to 625-Mc frequency band.

[1] Priest, D. H., *Proc. IRE*, **38**, 515, May, 1950.
[2] Radio Research Laboratory Staff, *op. cit.*, Chap. 19.

26. Barkhausen-Kurtz Oscillators. The extension of the frequency range of oscillators using conventional l-f circuits by the production of tubes with decreased electrode spacing seems to have reached a practical limit and oscillators based on entirely different principles have been responsible for major extensions of the frequency range into the millimeter wavelengths.

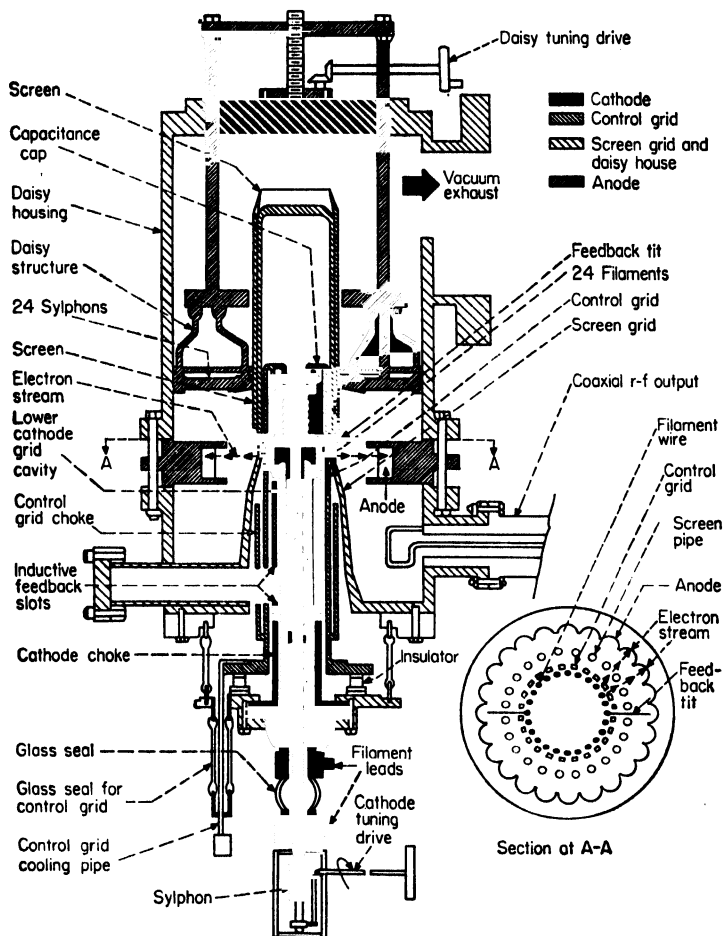



FIG. 34. Basic structure of Resnatron.

An early oscillator which made use of the transit-time effect for its operation was the Barkhausen-Kurtz oscillator,[1] which was developed in 1920. As shown in Fig. 35, this oscillator uses a fairly conventional triode operated with the plate at low negative potential and the grid at high positive potential. Electrons attracted from the

[1] Barkhausen, H., and K. Kurtz, *Physik. Z.*, **21,** 1, 1920. Gundlach, F. W., and W. Kleinsteuber, *Z. tech. Phys.*, **22,** 57, 1941. Kleinsteuber, W., *Hochfrequenztech.*, **53,** 199, 1939. Wundt, R., *Hochfrequenztech.*, **36,** 133, 1930. King, R., *Rev. Sci. Instruments*, **11,** 615, 1940. Hollman, H. E., "Physik und Technik der Ultrakurzen Wellen," Springer. Meinke, H., and F. W. Gundlach, "Taschenbuch der Hochfrequenztechnik," Springer, 1956.

cathode region by the positive grid penetrate the retarding field between the grid and plate and are stopped and returned to the grid. On the return trip, most of the electrons do not strike the grid but pass through into the grid-cathode region and then are reflected back again. Electron clouds therefore oscillate around the grid at a frequency primarily determined by the grid potential. Electrons which give up energy to the grid oscillate about the grid many more times than those which receive energy from the grid before striking it and, therefore, r-f oscillations can be maintained. A negative resistance consequently appears between the grid and plate, and if a resonator of sufficiently high impedance is connected across these electrodes, the circuit will oscillate. In general, the efficiency obtained with this type of oscillator is low and it has been replaced by two types of oscillators.

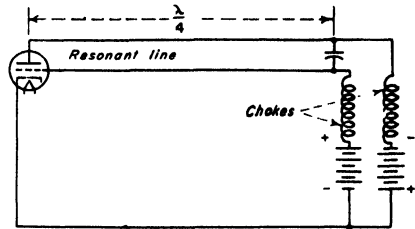



Fig. 35. Barkhausen-Kurtz oscillator employing resonant lines.

27. Klystron Oscillators. Klystrons[1] are h-f oscillators which depend upon transit time and velocity modulation of an electron beam for their operation and hence will oscillate at frequencies far above the maximum limit of triode oscillators. The word klystron is derived from the Greek "klyzein" which refers to the breaking of waves on a beach.

A klystron consists of a cathode, two cavity resonators, and a collector arranged as shown in Fig. 36. A cavity resonator[2] is a high-frequency resonant circuit in which the r-f electric and magnetic fields are completely contained within the resonator. Electrons are emitted from the cathode and travel toward the *buncher* gap under the influence of a high d-c voltage. The electrons pass through the buncher gap at a high velocity and then enter the field-free drift space in which their velocities remain constant. At the end of the drift space, the electron beam passes through the *catcher* gap and then strikes the collector. The grids on either side of the buncher and catcher gaps are connected to the cavity resonators tuned to the oscillating frequency. Therefore, if the buncher resonator is excited, an r-f voltage appears between the grids on either side of the buncher gap, which *velocity-modulates* the electron beam. This means that electrons which pass through the gap when the r-f electric field between the two grids is negative or retarding will be slowed down and electrons which pass through the gap when the r-f field is positive or accelerating are speeded up. Therefore, in the drift space, the accelerated electrons tend to catch up with the retarded electrons and form bunches as shown in the *Applegate diagram* in Fig. 37. As can be seen, the beam current passing through the catcher gap under these conditions is not uniform with respect to the time but is amplitude-modulated at the r-f frequency. The r-f component of the beam current passing through the catcher gap induces an r-f voltage across the catcher resonator and thus delivers r-f power to it. If some of the r-f energy from the catcher resonator is coupled back into the buncher resonator by means of a coupling loop or transmission line and the phasing is adjusted properly, oscillations will be maintained.

A fixed feedback circuit can be used and the phase adjusted for the desired oscillating conditions by means of the d-c voltage appearing between the cathode and buncher. This voltage controls the electron velocity and hence the time spent in the drift space which determines the phase of the electronic coupling between the buncher and catcher resonators. As shown in Fig. 38, oscillations will occur at several values of d-c voltage. Between the center of each mode, the electron drift time changes by the period of one r-f cycle. Oscillations will occur over a considerable range of voltage in each mode because the deviation in phase shift in the beam coupling from the desired value caused by the change in d-c

[1] Varian, R. A., and S. F. Varian, *J. Applied Phys.*, **10**, 321, May, 1939. Hamilton, D. R., J. K. Knipp, and J. B. H. Kuper, "Klystrons and Microwave Triodes," M.I.T. Radiation Laboratory Series, Vol. 7, McGraw-Hill, 1948. Spangenberg, K. R., "Vacuum Tubes," Chap. 17, McGraw-Hill, 1948. Reich, H. J., P. F. Ordung, H. C. Krauss, and J. C. Skalnik, "Microwave Theory and Techniques," Chap. 12, Van Nostrand, 1953. Harrison, A. E., "Klystron Tubes," McGraw-Hill, 1947.

[2] Hansen, W. W., *J. Applied Phys.*, **9**, 654, October, 1953. Hansen, W. W., and R. D. Richtmyer, *J. Applied Phys.*, **10**, 189, March, 1939.

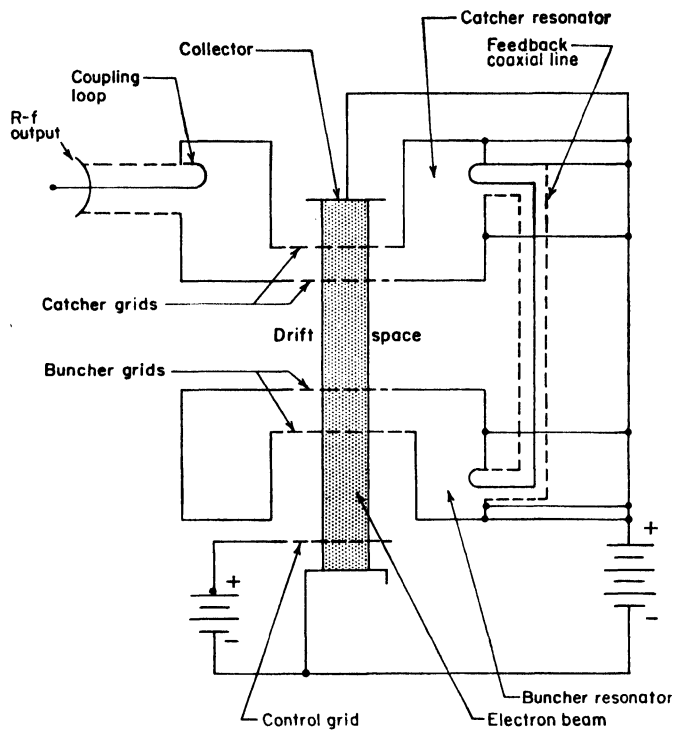



Fig. 36. Two-cavity klystron.

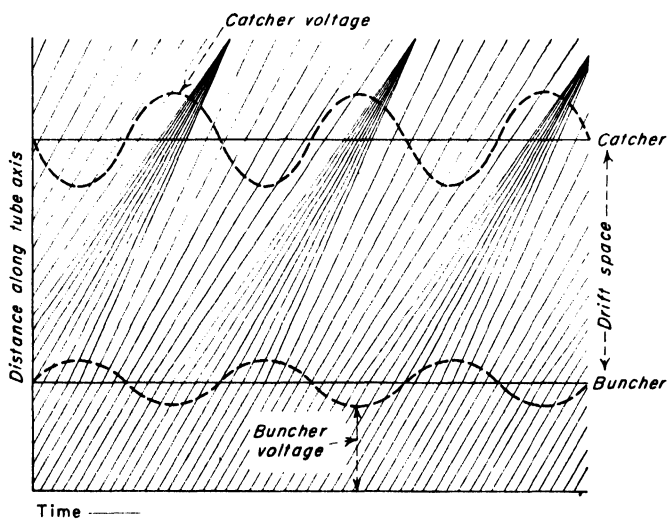



Fig. 37. Applegate diagram of two-cavity klystron.

voltage is compensated by a change in phase shift in the cavity resonators. The change in phase shift in the cavity resonators is produced by a small change in the frequency of oscillation. As a result, the oscillator can be frequency-modulated over ranges of the order of several megacycles by the application of a modulating voltage to the d-c beam voltage.

FIG. 38. Typical output power and frequency characteristics of type 410R two-cavity 3,000-Mc klystron oscillator.

Figure 39 shows a cross-sectional view of a typical two-cavity klystron. The resonators can be tuned over a relatively narrow frequency range by a variation in the spacing between the grids in each resonator which causes a change in the capacitance loading. One end of each resonator contains a flexible diaphragm which will allow axial movement of one of the grids without the use of complicated movable vacuumtight joints.

Klystrons have been developed which can be tuned over wide frequency ranges. In these tubes, the resonators are external to the vacuum and connections between the resonators and the grids at the buncher and catcher gaps are made by means of disks which pass through the vacuum seal.

Klystron oscillators are capable of producing large amounts of c-w or pulsed r-f power at microwave frequencies. However, for most high-power applications, three or more resonator klystron amplifiers are used rather than klystron oscillators because of the better frequency stability and higher efficiency obtained. Commercial klystron amplifiers are available which are capable of producing c-w power outputs of about 10 kw at frequencies up to 2,400 Mc and pulsed power outputs of 4 megawatts in the 3,000-Mc range. Efficiencies of from 30 to 45 per cent are obtained. Other experimental tubes, designed for pulsed operation, produce peak outputs of 30 megawatts at 3,000 Mc and still others generate power around the 150-watt level at frequencies up to 40 kMc.

FIG. 39. Structure of two-cavity klystron. (*Sperry Gyroscope.*)

28. Reflex-klystron Oscillators. A widely used type of microwave oscillator is the reflex-klystron oscillator.[1] This type of klystron requires only a single resonator,

[1] Pierce, J. R., *Proc. IRE*, **33**, 112, February, 1945. Hamilton, Knipp, and Kuper, *op. cit.* Ginzton, E. L., and A. E. Harrison, *Proc. IRE*, **34**, 97, 1946. McNally, J. O., and W. G. Shephard, *Proc. IRE*, **35**, 1424, 1947. Reich, Ordung, Krauss, and Skalnik, *op. cit.*, Chap. 13. Spangenberg, K. R., "Vacuum Tubes," McGraw-Hill, 1948.

as shown in Fig. 40, and hence is simpler in construction and operation than a multi-cavity klystron. The primary uses of this oscillator are in applications in which only small amounts of power are required, as for local oscillators in receivers, oscillators for signal generators, and low-power point-to-point microwave transmitters.

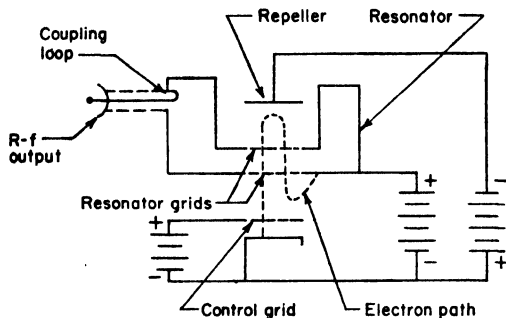



Fig. 40. Reflex klystron.

In a reflex klystron, the electron beam is accelerated by a d-c potential between the cathode and first grid of the resonator as in the two-cavity klystron. However, on leaving the resonator gap, the beam passes into a d-c retarding field space produced by a repeller electrode which is at a potential below that of the cathode. An electron will penetrate this space a distance which depends upon the velocity it has when entering the space. After being brought to rest by the retarding field, it will be accelerated in the opposite direction, thus passing back through the grids of the resonator. The same type of bunching action occurs in the reflex klystron as in the two-cavity tube. However, only one resonator is required and no external feedback connection is necessary. The proper phasing for oscillation is obtained by adjustment of the d-c repeller voltage. Figure 41 is a plot of the paths of various electrons in the retarding field area as a function of time under typical con-

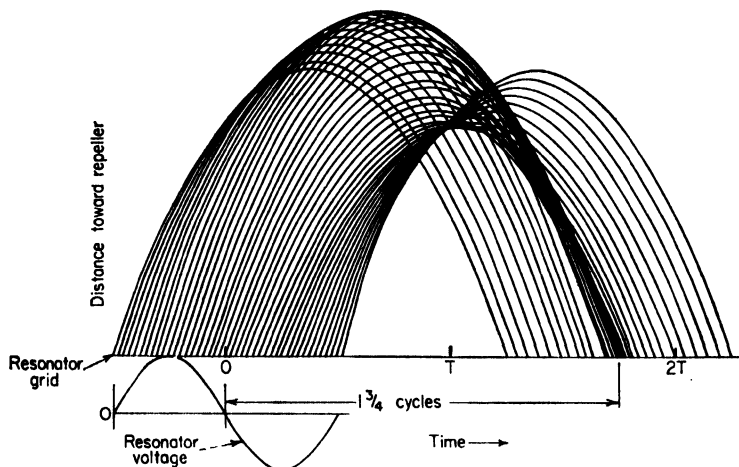



Fig. 41. Typical paths of electrons passing through the resonator at various times during an r-f cycle in a reflex klystron.

ditions and indicates how the bunches are formed. The variation in power output with repeller voltage is shown in Fig. 42; the several modes which can exist are indicated. As in the two-cavity klystron, the centers of the modes correspond to the transit angles of $(\frac{3}{4} + n)$ r-f cycles, where n is an integer called the *mode number*. The frequency of oscillations varies slightly over each mode as indicated. Consequently, the oscillator can be used to produce f-m signals or for automatic frequency control by the application of the modulating voltage to the repeller electrode. The highest-power mode of oscillation is the one in which the electrons require a time interval corresponding to three-quarters of an r-f cycle to make the round trip through the retarding-field space. However, this mode requires a very high resonator impedance for oscillation, and hence it is not often found. The power output and resonator impedance required generally drop as the mode number increases.

Reflex klystrons are commercially available with built-in resonators for frequencies between about 1,200 Mc and above 60 kMc with power outputs from a few milliwatts to several watts. Experimental tubes have been built which operate at considerably higher frequencies.[1] Tuning ranges of the order of 10 per cent are obtained. Figure 43 is a cross-section drawing of a typical narrow-band reflex klystron.

Wide tuning ranges can be obtained with reflex-klystron tubes designed for use with external resonators.[2] These tubes have the resonator grid connections brought out through the vacuum on copper disks as shown in Fig. 44. The external resonator can be a length of coaxial transmission line with a movable short circuit, as shown in Fig. 45, or a short-circuited wave guide.[3] The short circuit can be of the sliding-contact type or of the choke or noncontacting type (see Sec. 25). Tuning ranges as great as 5 to 1 have been obtained with the coaxial circuits at frequencies in the 1- to 10-kMc range and tuning ranges of the order of 2 to 1 are obtainable at even higher frequencies.

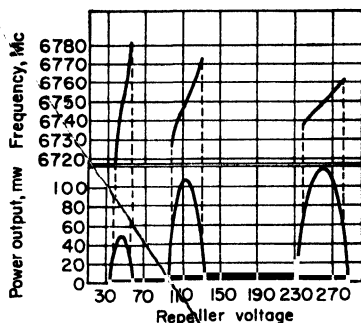



FIG. 42. Power output and frequency variation with repeller voltage for type 5976 reflex-klystron oscillator.

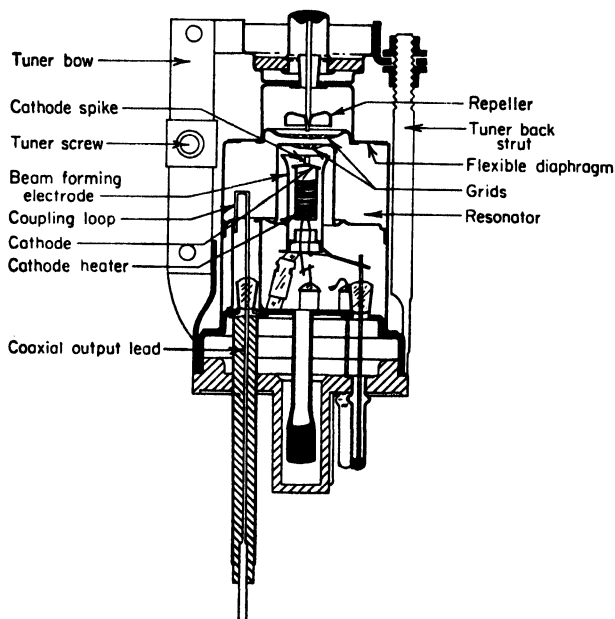



FIG. 43. Western Electric type 2K29 reflex klystron, 3,400 to 3,900 Mc.

If a single tuning control is desired, the repeller voltage must be tracked with the cavity tuning. The repeller voltage-frequency characteristic is practically linear;

[1] Reed, E. O., *Bell System Tech. J.*, **34**, 563, May, 1955.

[2] Radio Research Laboratory Staff, *op. cit.*

[3] Myers, W. D., *Hewlett-Packard J.*, **2**, September, 1950. Fong, A., *Hewlett-Packard J.*, **5**, January-February, 1954. Harman, W. W., *Proc. IRE*, **38**, 671, June, 1950.

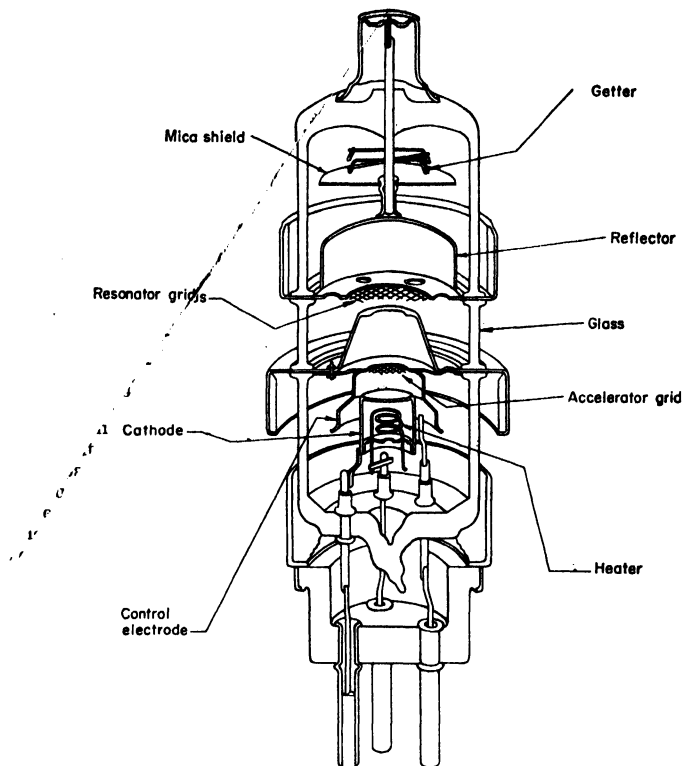



Fig. 44. Sylvania type 6BL6 reflex klystron for use with external resonator.

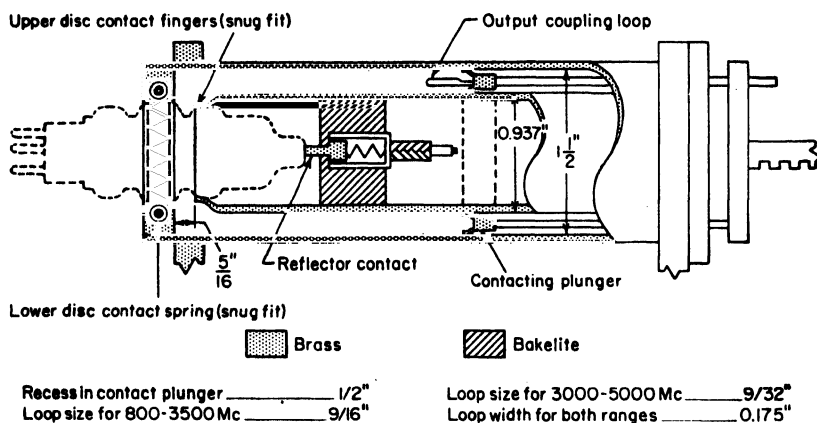



Fig. 45. Coaxial cavity for a wide-range reflex klystron using type 6BM6A tube for 800- to 3,500-Mc range and type 6BL6A tube for 1,200 to 5,000 Mc.

however, the plunger position in the cavity is approximately proportional to wavelength rather than to frequency. Many methods,[1] such as the use of nonlinear gears, a tapped voltage-divider system, a tapered voltage-divider system, or a variable-transformer system, are satisfactory for tracking purposes.

29. Magnetron Oscillators. An important type of h-f oscillator is the magnetron.[2] Magnetrons have been very widely used for the generation of pulsed r-f power in radar systems at microwave frequencies. Efficiencies up to 80 per cent and peak power outputs as high as 10 megawatts have been obtained in the microwave frequency band. Recently, a new type of magnetron has been developed which can be tuned by variations in voltage alone over frequency ranges as large as 4 to 1.

In a magnetron, electrons move under the influence of crossed electric and magnetic fields. In general, a magnetron consists of a cylindrical cathode surrounded by a cylindrical anode. A d-c voltage is applied between the cathode and anode, and a magnetic field is applied parallel to the axis of the cylindrical electrodes. An electron which leaves the cathode under the influence of the d-c voltage is deflected from a radial path between the cathode and anode by the force produced by the magnetic field[3] as shown in Fig. 46. This

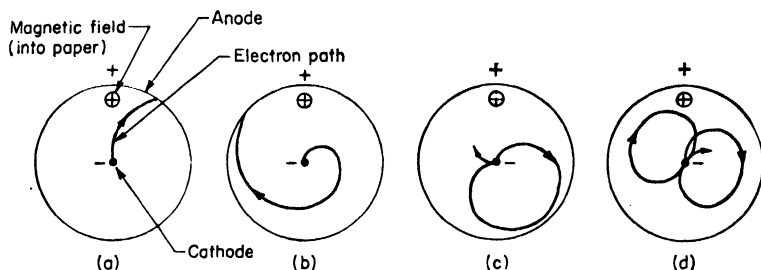



Fig. 46. Typical electron paths in cylindrical magnetron. The potential between cathode and anode is the same in all cases, but the magnetic field increases from (*a*) to (*b*).

force is proportional to the strength of the magnetic field and to the velocity of the electron. Its direction is at right angles to both the magnetic field and the direction of motion of the electron. Therefore, with proper settings of the magnetic field and the d-c voltage, electrons emitted from the cathode can be made to travel directly from the cathode to the anode, to spiral around the cathode before striking the anode, or to loop around and return to the cathode, in which case the anode current is cut off. The voltage which marks the transition between a finite anode current and zero anode current with a fixed value of magnetic field is called the *cutoff voltage*. Theoretically under the proper conditions, electrons can be made to circle about the cathode indefinitely and never return to either the anode or cathode. However, it is not believed that this condition is obtained in practice. Most magnetrons are designed to be in the cutoff region when the r-f field is absent.

There are three basic types of magnetrons. The *negative-resistance magnetron* depends for its operation upon a static negative resistance produced under certain conditions in a magnetron. This type of operation is affected by transit time and hence is usually useful only at the lower frequencies. The *cyclotron-frequency magnetron* depends for its operation upon the periodic radial component of motion of the electrons parallel to and in synchronism with a radial r-f electric field. In the *traveling-wave magnetron*, the r-f field is primarily a tangential field and produces bunching of the electrons around the interaction space like spokes on a wheel. This space-charge cloud rotates in synchronism with the rotating traveling r-f field which is set

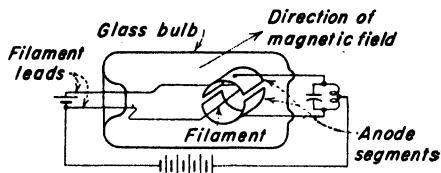
[1] Radio Research Laboratory Staff, *op. cit.*, Sec. 31-12.

[2] Fisk, J. B., H. D. Hagstrum, and P. L. Hartman, *Bell System Tech. J.*, **25**, April, 1946. Also Bell Laboratory Staff, "Radar Systems and Components," Van Nostrand, 1949. Collins, G. B., "Microwave Magnetrons," M.I.T. Radiation Laboratory Series, Vol. 6, McGraw-Hill, 1948. Spangenberg, K., "Vacuum Tubes," McGraw-Hill, 1948. Slater, J. C., "Microwave Electronics," Chap. 13, Van Nostrand, 1950. Reich, Ordung, Krauss, and Skalnik, *op. cit.*, Chap. 14.

[3] Hull, A. W., *Phys. Rev.*, **18**, 31, July, 1921. Brillouin, L. N., *Phys. Rev.*, **60**, 385, Sept. 1, 1941.

up in interaction space between the anode and cathode in the magnetron. Most high-power modern magnetrons are of the latter type.

30. Negative-resistance Magnetron Oscillators. Negative-resistance magnetron oscillators are usually of the split-anode type shown in Fig. 47 in which the cylindrical anode is split into two equal segments. The frequency of oscillation is determined by the resonant frequency of the tuned circuit connected between the two anode segments and the operation is not critically dependent on the magnetic field.

FIG. 47. Split-anode magnetron.

In this oscillator, the d-c voltage between the anodes and the cathode is well below the cutoff voltage. When, in addition, an a-c voltage is applied across the anode segments, the potentials of the two anode segments are periodically raised and lowered in opposite phases. Under these conditions, electrons tend to spiral out from the cathode and strike the anode segment at the lowest potential, as shown in Fig. 48, thus producing a negative a-c resistance.[1] Negative-resistance magnetrons are limited to relatively low-frequency operation since the transit time of electrons between the cathode and anode must be a small fraction of an r-f cycle. The upper frequency limit is of the order of 1,500 Mc, and in this range, very large magnetic fields are required.

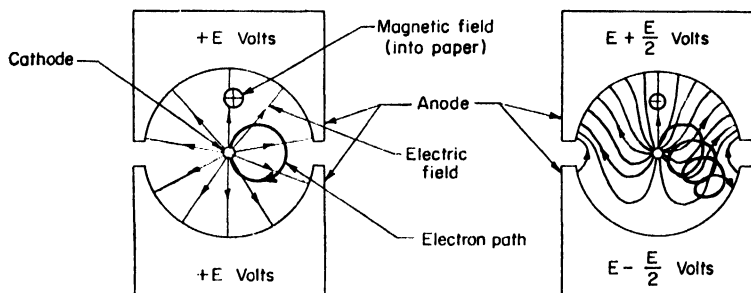



FIG. 48. Typical electron paths in negative-resistance split-anode magnetron.

31. Cyclotron-frequency Magnetron Oscillators. An electron which leaves the cathode under the influence of the anode voltage will start toward the anode. However, as it gains velocity, it will be deflected by the right-angle force produced by the magnetic field which is proportional to the velocity of the electron and to the strength of the magnetic field. If the anode voltage is below the cutoff value, the electron will loop around and return to the cathode. The frequency corresponding to the time required for an electron to complete one loop or cycle in a parallel-plane magnetron is called the cyclotron frequency f_c. The cyclotron frequency in a parallel-plane magnetron is determined only by the strength of the magnetic field:

$$f_c = 2.8 \times 10^6 B$$

where f_c is the frequency in cycles and B is the magnetic field strength in gausses. In a cylindrical magnetron, the operation is similar, as shown in Fig. 46, but the frequency of the radial component of the electron motion is lower than the cyclotron frequency.

If an r-f voltage having a frequency equal to that of the radial component of the electron motion is applied between the cathode and anode, electrons which leave the cathode in such a phase as to gain energy from the r-f field as they travel toward the anode will also gain

[1] Kilgore, G. R., *Proc. IRE*, **24**, 1140, 1936.

energy from the r-f field on their return to the cathode. These electrons will therefore strike the cathode with more energy than they had when they left and hence will be removed from further interaction with the r-f field. However, electrons which leave the cathode in such a phase as to give up energy to the r-f field as they move toward the anode will also lose energy on their return to the cathode. These electrons will not have sufficient energy to reach the cathode and will therefore start back toward the anode on another loop and give up still more of their energy to the r-f field. These electrons will continue to give up energy to the r-f field until the amplitude of their radial oscillation becomes zero. At this time, they must be removed or they will start to absorb r-f energy. In general, electrons emitted from the cathode which tend to absorb r-f power are rapidly eliminated and electrons which deliver power to the r-f field are allowed to multiply. As a result, if a resonant circuit is connected between the cathode and anode and tuned to the frequency of the radial oscillation of the electrons, r-f power will be generated.

Cyclotron-frequency magnetrons can also be of the split-anode type, but neither type is very widely used because the frequency is critically dependent on the magnetic field and both the power output and efficiency are low. Power outputs of 100 watts can be obtained at a frequency of 600 Mc and 1 watt at 3,000 Mc. Efficiencies lie between 1 and 10 per cent.

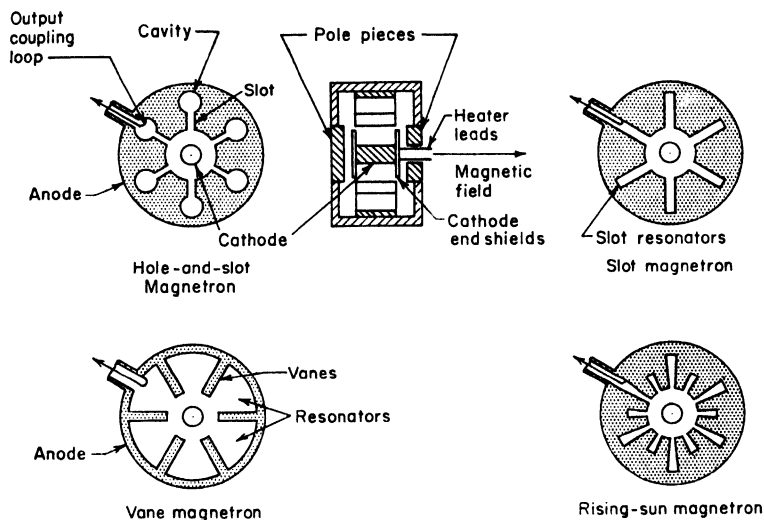



FIG. 49. Cross sections of several types of multicavity magnetrons.

32. Traveling-wave Magnetron Oscillators. Traveling-wave magnetron oscillators are the most widely used type of magnetron because of their high efficiency, high power output, and adaptation to use over frequency ranges of from a few hundred megacycles to the millimeter range. The operation of this type of magnetron depends upon the interaction of a rotating r-f field with a synchronously rotating space-charge cloud in a manner analogous to an a-c generator.

The traveling r-f wave can be set up by a number of different types of structures, and magnetrons are usually classified by the structure used. These classifications are:

1. Multicavity magnetrons
2. Split-anode magnetrons
3. Interdigital magnetrons

Multicavity magnetrons are built in several forms, some of which are: (1) hole-and-slot magnetron, (2) slot magnetron, (3) vane magnetron, and (4) rising-sun magnetron. Cross-sectional views are shown in Fig. 49. In all these magnetrons, a number of

cavities are located around the circumference of the anode. The cavities are parallel-resonant circuits which are all connected in series in a ring as shown in the equivalent circuit in Fig. 50. If this circuit is excited by r-f energy at the proper frequency, a

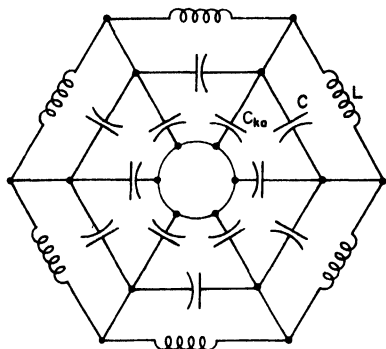



Fig. 50. Approximate equivalent of a hole-and-slot magnetron.

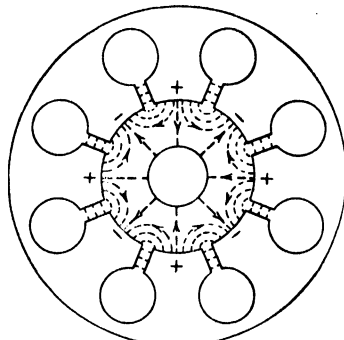



Fig. 51. Electric field distribution in a hole-and-slot magnetron operating in the π mode.

standing wave will be set up around its circumference. A standing wave is equivalent to two waves of equal amplitude traveling in opposite directions. The wave traveling in the same direction and at the same speed as the rotating space charge interacts with it, while the wave traveling in the opposite direction has little effect. Standing waves can be set up at a number of frequencies corresponding to different modes of operation, a characteristic which can be very troublesome. Usually magnetrons are designed to operate in the π mode, in which adjacent cavities are 180 deg out of phase with one another as indicated in Fig. 51. Low-frequency magnetrons are usually "strapped" to force them to operate in the π mode. Strapping consists in connecting together alternate anode segments, which are in the same phase in the π mode, by means of copper straps, as shown in Fig. 52.

Fig. 52. Structure of a strapped magnetron. (*From Slater.*)

In a traveling-wave magnetron, electrons under the influence of the axial magnetic field and radial electric field follow approximately epicycloidal paths which progress around the cathode. The average velocity of the progression depends on the relative strengths of the d-c electric and magnetic fields. If the fields are chosen properly, the average rotational velocity can be made equal to that of the traveling r-f wave propagated along the anode. In a magnetron operating in the π mode, the r-f electric field is as indicated in Fig. 51. An electron which encounters .an accelerating tangential component of the r-f electric field absorbs energy from the field and has its velocity increased and thus tends to return to the cathode. An electron which encounters a retarding tangential component of the r-f field delivers energy to the field and tends to have its velocity reduced. The lower velocity

causes this electron to move closer to the anode, and in doing so, it obtains more energy from the d-c electric field and therefore tends to regain its original rotational velocity. These electrons therefore remain in synchronism with the rotating r-f electric field and continue to abstract energy from the d-c field and deliver energy to the r-f field until they strike the anode. Figure 53 shows typical paths of electrons and indicates the space-charge bunches which rotate about the interaction space like spokes on a wheel. The radial

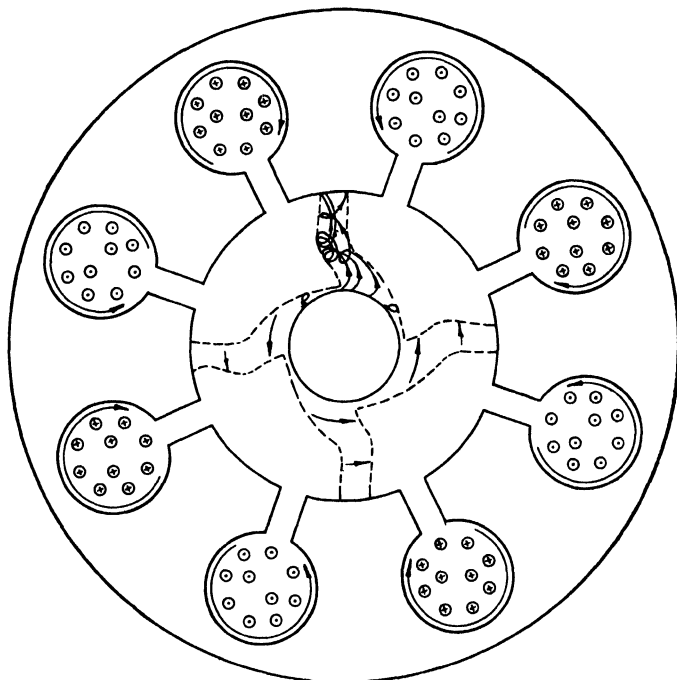



FIG. 53. Space-charge distribution and electron paths in magnetron. The paths indicated are relative to the space-charge cloud which is rotating in synchronism with the traveling r-f wave.

component of the r-f field also helps to maintain the electrons in the space-charge bunches in the proper phase by a process known as *phase focusing*.

Split-anode magnetrons operating in the traveling-wave mode are similar to multicavity magnetrons in operation, with the difference that the rotational velocity of the traveling r-f wave is much larger since there are only two anode segments. This requirement tends to limit the maximum frequency of oscillation with practical values of d-c voltage and magnetic field.

Interdigital magnetrons are traveling-wave magnetrons utilizing a single resonator in which the anode segments are fingers extending from both ends of the resonator, as shown in Figs. 54 and 55. Adjacent fingers are connected to opposite ends of the resonator and hence are 180 deg out of phase. A rotating traveling wave is set up in the same manner as in multicavity magnetrons. Since only a single resonant circuit is required, these magnetrons can be easily adapted to being tuned over wide frequency ranges.

The performance of magnetron oscillators is greatly affected by the load impedance appearing at the coupling loop. The actual power output and frequency of oscillation

for various load impedances can be determined from a Rieke diagram, shown in Fig. 56. In this diagram, contours of constant power and contours of constant frequency are plotted on a Smith chart in which constant-VSWR or reflection-coefficient contours are circles about the center of the chart and contours of constant distance along the

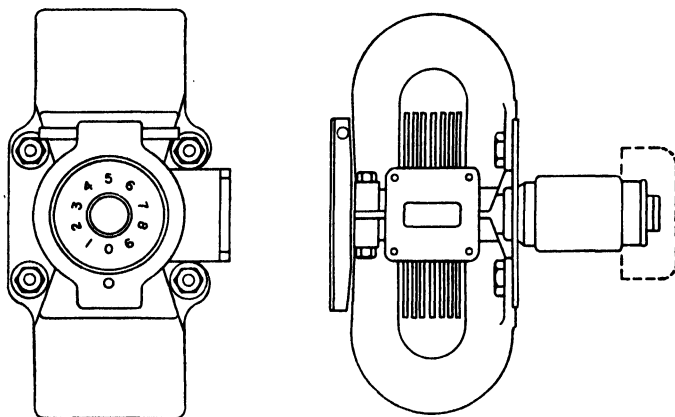



Fig. 54. Typical traveling-wave magnetron with magnet. (*From Sylvania.*)

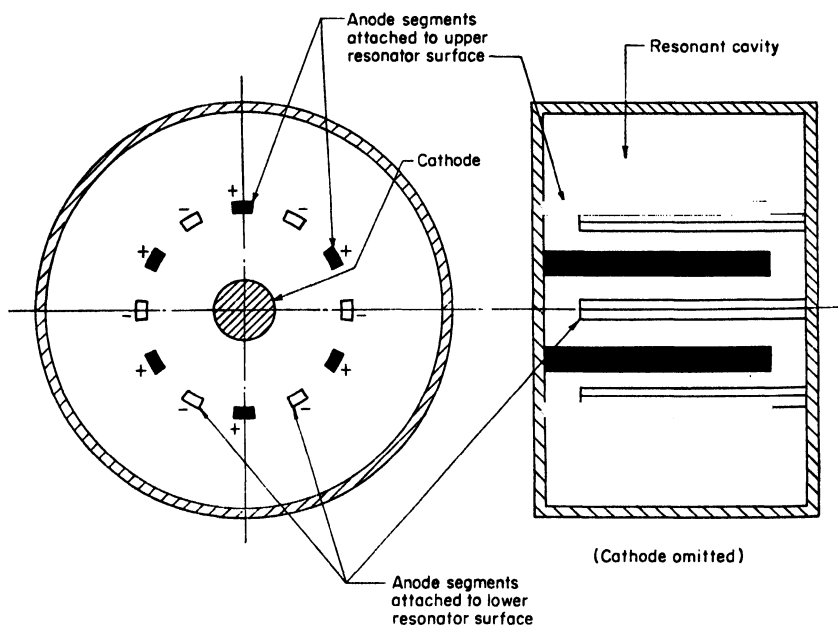



Fig. 55. Interdigital magnetron construction.

line from a reference point, usually the output terminals, to a voltage minimum on the output line or wave guide are radial lines.

33. Tunable Magnetron Oscillators. As pointed out above, interdigital and split-anode magnetrons can be tuned over wide frequency ranges by adjustments

of the resonator tuning. Multicavity magnetrons are more difficult to tune but tuning over about 10 per cent ranges can be obtained by the insertion of rods in the cylindrical cavities or plates in the slots of hole-and-slot magnetrons. The rod method is mainly inductive tuning while the plate method is mainly capacitive tuning.

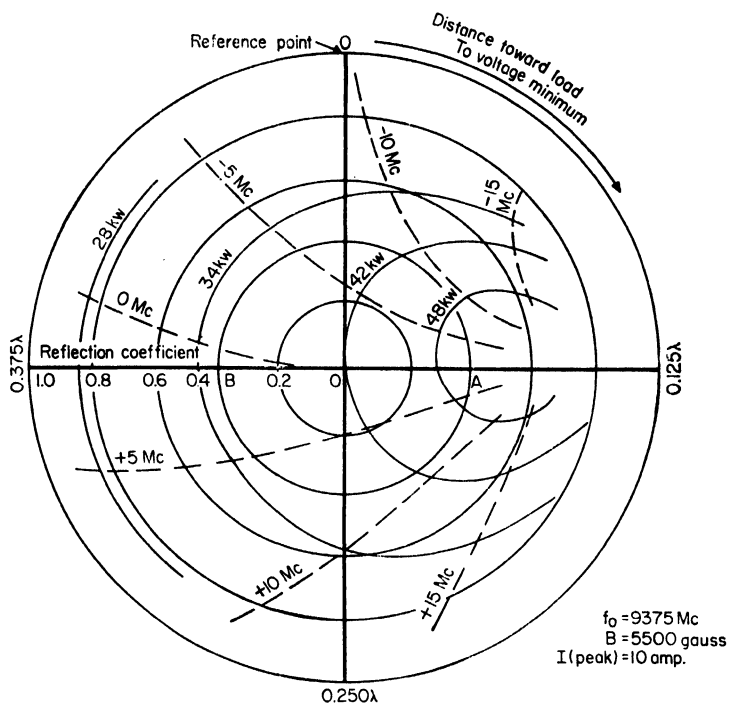



FIG. 56. Typical Rieke diagram.

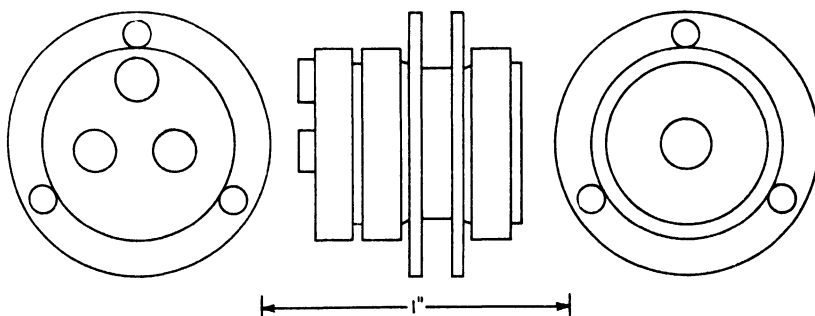



FIG. 57. Type Z5112 voltage-tunable magnetron. (*From General Electric.*)

34. Voltage-tunable Magnetron Oscillators. Interdigital magnetrons can be constructed which can be tuned by variations in voltage alone over frequency ranges as great as 4 to 1. An outline view of one version of this type of tube[1] is shown in Fig. 57.

[1] Welch, H. W., Jr., *Proc. IRE*, **41**, 1631, November, 1953. Boyd, J. A., *Proc. IRE*, **43**, 332, March, 1955. Wilbur, D. A., and P. H. Peters, *Proc. Nat. Elec. Conf.*, **11**, 368, 1955.

As previously pointed out, in a traveling-wave magnetron the traveling wave must stay in synchronism with the rotating space charge if oscillations are to be produced. Therefore, the velocity of the traveling wave and hence the frequency of oscillation are proportional to the rotational velocity of the space-charge wave. Since the velocity of the space-charge wave is proportional to the applied d-c voltage, the oscillating frequency also will be a linear function of the applied voltage.

Most magnetrons cannot be voltage-tuned over wide frequency ranges because the Qs of the associated anode resonators are too high, and if the Qs are lowered by loading, the oscillations cease before the tuning range can be greatly increased. It has been found, however, that if the cathode current is maintained constant, the oscillator will function with load impedances of only a few ohms at a frequency primarily determined by the applied d-c voltage. Under these conditions, the magnetron

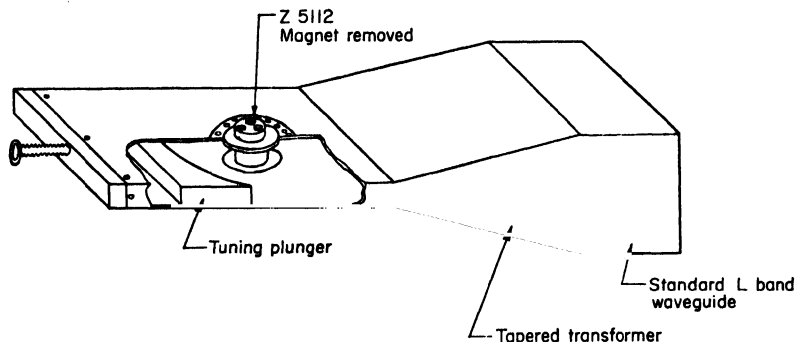



FIG. 58. Tapered wave-guide cavity for voltage-tunable magnetron.

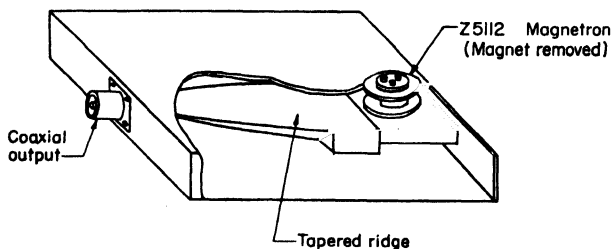



FIG. 59. Ridge wave-guide cavity for voltage-tunable magnetron.

appears to be a constant-current generator whose output terminals are shunted by the capacitance between the vanes of the interdigital structure. The power output is therefore approximately proportional to the load resistance. Since the load is shunted by the fixed vane capacitance, the band width of the output circuit and the tuning range are inversely proportional to the load resistance and the power output. Tuning ranges of over 4 to 1 have been obtained with a power output of about ¼ watt. For a 2:1 band width, a power output of about 1 watt is obtainable, and over band widths of a few hundred megacycles, an output of 10 watts can be obtained.

Since the required load impedance for broad-band operation is in the vicinity of 10 to 20 ohms, broad-band matching transformers are necessary to match the tube to a 50-ohm load or to a standard wave guide. Figure 58 shows a tapered wave-guide form of matching transformer used in the 2.5- to 4.0-kMc range, and Fig. 59 shows a ridge wave-guide cavity for the 3.5- to 4.7-kMc range. The power output as a function of frequency of the tapered wave-guide oscillator is shown in Fig. 60.

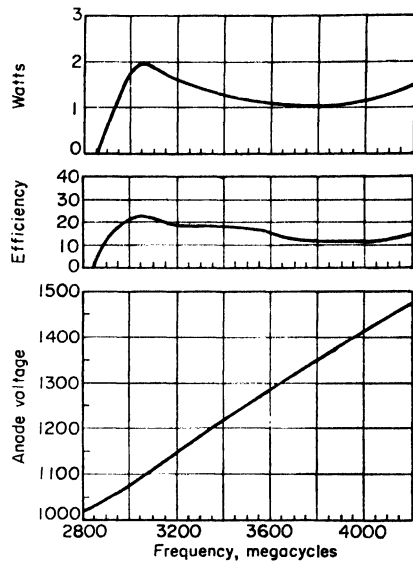
One problem in these tubes is that of maintaining constant cathode current. This can be accomplished by operating the cathode temperature-limited. However, back heating of the cathode by the electrons which are returned to the cathode, after having gained energy from the r-f field, makes it necessary to control the heater power supplied to maintain a constant cathode temperature. A method of overcoming this difficulty has been found. The technique is to locate the actual emitting cathode outside of the actual interaction space and to deflect the electrons into the interaction space by means of a control electrode, as shown in Fig. 57.

The plate voltage required to operate the tube is of the order of 500 to 1,500 volts and the magnetic field about 2,500 gausses. The voltage sensitivity is about 3 Mc per volt. For best operation, the tube should have a regulated d-c heater supply. The noise level in the tube has been found to be slightly higher than that of klystrons and triodes.

At present (1958), tubes are commercially available to cover the frequency range 1,000 to 4,000 Mc (Z5111, 1,000 to 2,000 Mc, and Z5112, 2,000 to 4,000 Mc) and other models are planned to extend this range in both directions. This type of oscillator is very useful for microwave sweep generators, spectrum analyzers, wide-range receivers, and various other low-power applications.

FIG. 60. Performance of type Z5112 voltage-tunable magnetron.

35. Backward-wave Oscillators.[1] Another new type of microwave oscillator is the backward-wave oscillator, some versions of which are called *Carcinotrons*. This oscillator has the important characteristic of having its frequency of oscillation controllable over ranges greater than 2 to 1 by voltage variation alone. This device is similar to a traveling-wave-tube amplifier (see Chap. 8) but differs in that the electron beam travels in the opposite direction from the wave in the slow-wave structure and as a result produces the feedback required for oscillation.

In a backward-wave oscillator, the electromagnetic wave is propagated in a slow-wave structure in which the effective velocity of the wave interacting with the electron beam has its phase velocity and group velocity in opposite directions. In Fig. 61, the slow-wave structure consists of a folded wave guide which intersects the beam periodically. If an electromagnetic wave is applied to the wave guide from the right-hand end, it will travel toward the left-hand end and interact with the beam at the points at which the guide intersects the beam. If the frequency of the wave and the electron velocity are adjusted so that an electron which encounters a peak retarding field when it passes through the extreme left-hand interaction gap also encounters a peak retarding field when it passes the next gap and all subsequent gaps, the beam will be bunched as it travels along the line and will deliver power to the wave in the same manner as in the traveling-wave tube. Note that, although the electromagnetic wave in the guide is traveling to the left, it appears to the electron beam that it is traveling to the right in synchronism with the beam. The beam delivers power to the r-f wave as it travels to the right and hence actually generates the electromagnetic wave in the guide which causes the bunching. The feedback loop con-

[1] Kompfner, R., and N. Williams, *Proc. IRE*, **41**, 1602, November, 1953. Heffner, H., *Proc. IRE*, **42**, 936, June, 1954. Johnson, H. R., *Proc. IRE*, **43**, 684, June, 1955. Warnecke, R. R., P. Guenard, O. Doehler, and B. Epstein, *Proc. IRE*, **43**, 413, April, 1955. Palluel, P., and A. K. Goldberger, *Proc. IRE*, **44**, 333, March, 1956.

sisting of the electromagnetic wave and electron beam produces self-sustaining oscillations if the beam current is sufficiently large, and r-f power output can be obtained from the left end of the wave guide. The right end of the guide is usually terminated to prevent reflections from interfering with the action.

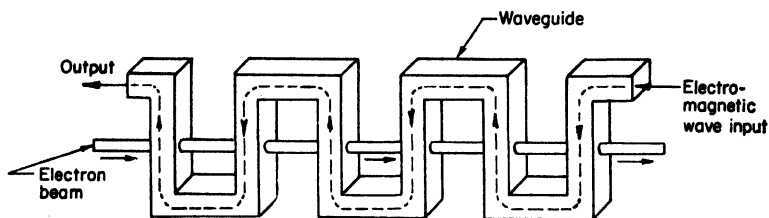



FIG. 61. A basic type of backward-wave oscillator. (*After Pierce.*)

Since the electron beam actually generates the wave in the wave guide, the frequency of the wave will automatically adjust itself to that providing an effective phase velocity which produces the optimum oscillating conditions. As a result, the operating frequency can be varied over wide ranges by variations in the beam velocity produced by changes in the d-c beam voltage.

In practice, the interaction structure may be a helix, folded line, interdigital structure, or any other periodic structure and some types of smooth structures.

36. O-type Backward-wave Oscillator. There are two basic types of backward-wave oscillators, the O-type and M-type. The O-type oscillator is similar to a conventional traveling-wave tube in that the electrons are accelerated by a d-c electric field before entering the interaction structure only, with no d-c electric fields present in the interaction space, and the r-f energy is obtained from the kinetic energy of the electron motion.

Figure 62 is a diagram of an O-type backward-wave oscillator. A magnetic field is used for focusing purposes only and is oriented parallel to the beam. The efficiency of the

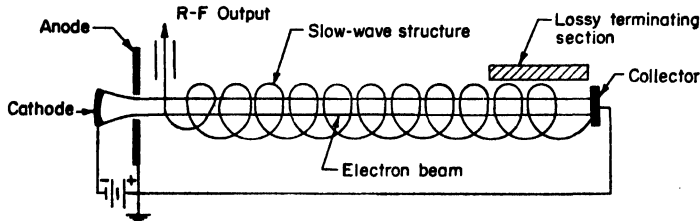



FIG. 62. Structure of O-type backward-wave oscillator with a helical slow-wave structure.

O-type oscillators is usually low, in the vicinity of 1 to 5 per cent, since the electron beam is slowed down as it gives up energy and can reach a velocity at which it absorbs energy from the r-f circuit. This type of oscillator has a voltage-frequency characteristic which is roughly proportional to f^3. The noise produced is comparable with that produced by a klystron. O-type backward-wave oscillators have been built which are voltage-tunable over about an octave in frequency with center frequencies from a few hundred megacycles up to frequencies in the millimeter range. Power outputs range from several watts at the lower frequencies to a few milliwatts at millimeter wavelengths. In general, this type of tube is suited to local oscillator and instrument applications.

37. M-type Oscillator. In the M-type backward-wave oscillator,[1] the beam is acted on by crossed electric and magnetic fields in the interaction area, as shown in

[1] Palluel and Goldberger, *loc. cit.*

Fig. 63, and the r-f energy is obtained from the potential energy of the electrons rather than from the kinetic energy. This type of oscillator is similar in principle to a magnetron.

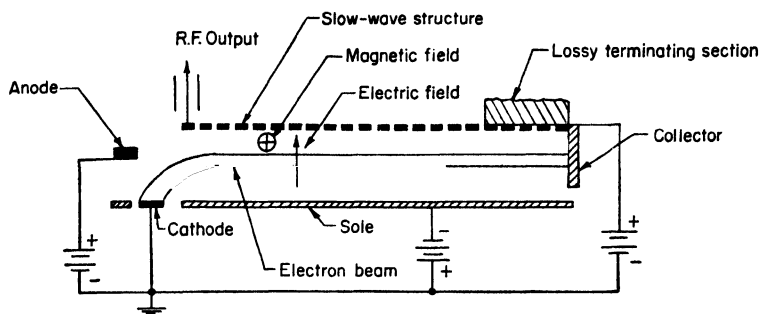



FIG. 63. Structure of M-type backward-wave oscillator.

The electron beam is introduced at right angles to the axis of the periodic structure and under the influence of the magnetic field initially follows a curved path. At a point at which the direction of motion of the beam is parallel to the axis of the periodic structure, it is injected into the interaction region, where it also is acted on by the d-c electric field existing between the periodic structure and the sole. Under the influence of crossed electric and magnetic fields, the beam will follow a path parallel to the axis of the periodic structure. The reaction with the periodic structure is similar to that in the O-type oscillator. However, as the electrons give up energy to the r-f circuit, they do not slow down but move to a region of higher potential energy in the transverse d-c electric field. The efficiency of this type of oscillator is high. Theoretical efficiencies of 80 per cent have been predicted and in practice efficiencies of 20 to 40 per cent have been obtained. Tubes have been built with power outputs in the kilowatt range. The voltage-frequency characteristic of the M-type backward-wave oscillators is linear.

RC OSCILLATORS

38. Fundamentals. These oscillators differ in principle from *LC* oscillators in that an *RC* network replaces the tuned *LC* circuit as the coupling device between the output of the amplifier and its input. In certain *RC* oscillators, the performance of the network parallels that of the *LC* tuned circuit, and Q can be defined in a manner identical to that for tuned circuits. In other *RC* circuits no Q can be defined on the basis of any of the three accepted definitions for Q of an *LC* circuit. Brown[1] has proposed that $Q = \frac{1}{2}[d\phi/d\alpha]_{\alpha=1}$ be used, that is, as one-half the rate of change of phase with the fractional change of frequency at the operating point $\alpha = \omega/\omega_0 = 1$. By this definition values of Q range below and above unity. Morris[2] shows that defining Q in this manner leads to different values of Q depending upon the method of use of the *RC* network, *i.e.*, upon the terminal pair chosen for the output. He then proposes a mathematical concept, based on the differential equation of the circuit. This is most easily understood by an example. In Fig. 64

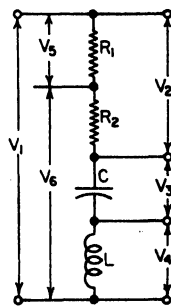



FIG. 64. Definition of Q as proposed by Morris.

$$\frac{V_2}{V_1} = \frac{R}{1/pC + R + pL}$$

where $R = R_1 + R_2$
$p = d(\quad)/dt$

[1] Brown, D. A. H., The Equivalent Q of RC Networks, *Electronic Eng.*, July, 1953, p. 294. Francini, Giuseppe, Evaluation of Oscillator Quality, *IRE Trans. Circuit Theory*, CT-2, No. 3, September, 1953.
[2] Morris, David, Q As a Mathematical Parameter, *Electronic Eng.*, July, 1954, p. 306.

Write the denominator in ascending powers of (pT), making the first term unity. Choose the time constant T so that the coefficient of $(pT)^2$ is unity. The reciprocal of the coefficient of (pT) is defined as Q. Then

$$\frac{V_2}{V_1} = \frac{pT/Q}{1 + pT/Q + (pT)^2}$$

where $T = \sqrt{LC}$
$Q = (1/R)\sqrt{L/C}$

If the similar ratios are formed for the several outputs indicated, the numerators of the expressions differ but the denominators remain the same. Defined in this way, Q is characteristic of the given assembly of components rather than of the particular

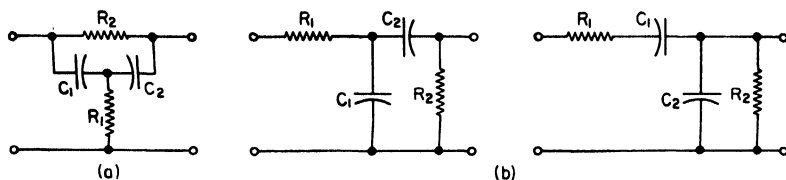



FIG. 65. If the circuit in (a) is rearranged as in (b), Q can be defined in conventional manner.

output that is taken from the components. Furthermore, the definition of Q can be extended to uses for which the conventional definitions are inapplicable.

For example, in the bridged T of Fig. 65a, Q cannot be defined according to any one of the three conventional definitions of Q. Following Morris's proposal Q can be defined.

The same components arranged as in Fig. 65b yield circuits in which Q can be defined in the conventional manner, $Q = f_0/(f_2 - f_1)$, where f_0 = resonant frequency and f_1, f_2 = frequencies at half-power points.

39. *CR* and *RC* Ladders. These are illustrated in Fig. 66a and b. For three sections up to six sections, the loss in either type is

No. Sections	Loss
3	1/29
4	1/18.4
5	1/15.4
6	1/14.1

The gain required in an amplifier to obtain oscillation is 1/loss.

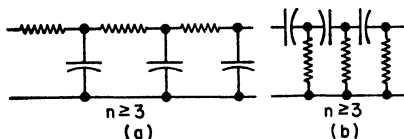



FIG. 66. *CR* and *RC* ladders.

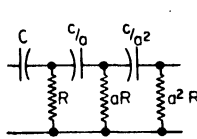



FIG. 67. *RC* ladder with graded sections.

Some improvement is obtained in the required gain if the successive sections are made of higher and higher impedance, as shown in Fig. 67, for three sections:

Multiplier a	Loss
1	1/29
2	1/16
3	1/12.8
4	1/11.4

40. Lead-Lag or Advance-Retard Networks. Little or no mention is made of "advance-retard" or "retard-advance" ladders in the literature. Papers show four

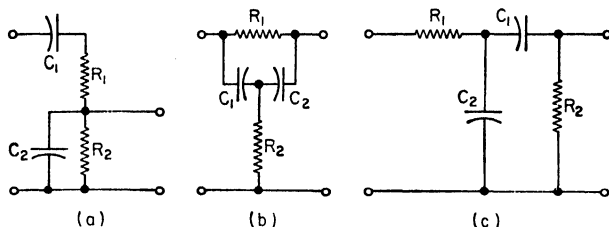



FIG. 68. The same frequency response is obtained from a Wien-bridge chain (*a*), a bridged-T (*b*), or a *CR* section followed by an *RC* section (*c*) or vice versa.

sections of "advance" (*CR*) or "retard" (*RC*) ladders, for example, with the idea that a total phase change of ±180 deg must be obtained. Many times, a net phase change of zero, produced by a ±90 deg shift followed by a ∓90 deg shift, leads to more desirable results.

The same performance as regards frequency response is obtained from the four elements of a *Wien-bridge* chain or of a bridged-T circuit by using them in a *CR* section followed by an *RC* section, or vice versa, as shown in Fig. 68.

41. Wien-bridge Oscillators. The Wien-bridge circuit is often used in *RC* oscillators, either in the series-parallel chain or with the chain as one half of a bridge with two resistors as the opposite half. Both are shown in Fig. 69*a* and *b*.

FIG. 69. Wien-bridge chain and a bridge form commonly used in *RC* oscillators.

Davidson[1] points out that, if the compensating resistor R is not used, the frequency of oscillation is

$$\omega_0 = \frac{1}{RC\sqrt{1 + R_A/R}}$$

where R_A is the source impedance.

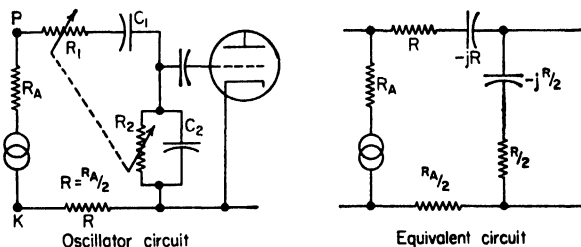



FIG. 70. Compensation of an *RC* oscillator to make $\omega_0 = 1/RC$.

If the compensating resistor R is added (Fig. 70) and made equal to $R_A/2$, the frequency is $\omega_0 = 1/RC$ and precision decade resistors can be used to predetermine the

[1] Davidson, J. A. B., A Note on Precision Decade Oscillators, *Proc. IRE*, Correspondence, September, 1952, p. 1124.

frequency. At the higher frequencies, if R is slightly greater than $R_A/2$, some compensation for phase shift in the amplifier can be obtained.

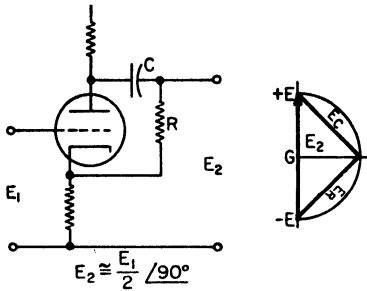



Fig. 71. Phase-shifting circuit giving 90° phase shift.

42. Phase Inverters with *RC* or *CR* Networks. By use of a phase-inverter tube, an *RC* network which normally gives only ±45-deg phase shift can be made to give 90-deg shift. Two successive such stages provide 180-deg shift with a loss of only 4. The circuit is shown in Fig. 71, with the vector diagram. An oscillator based on this phase-shifting circuit is described by Frazer.[1] Rosenstreter[2] described it as part of "more complex apparatus" and Villard[3] described it as a tunable amplifier.

This idea is easily applied in making a source of voltage adjustable in phase, as described by Rao.[4]

A very good amplitude regulator, good from audio frequencies to 10 Mc or more, is another adaptation of the phase inverter, as shown in Fig. 72.

With *LC* and *RC* oscillators, where the driving amplifier has a large excess of gain, the use of this control will maintain the level practically constant. Using 6-watt 115-volt tungsten lamps, a balance will be obtained on the bridge, with about 1 volt on each half of the bridge, at which point the lamp resistances are about 370 ohms, which is the value of the fixed arms.

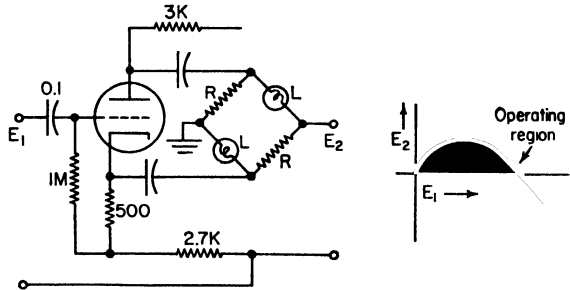



Fig. 72. Automatic amplitude regulator.

This form of *RC* phase shifter, applied in an oscillator, provides a feature which has been described only by Smiley and is given in a short summary below. If a voltage divider is used, a voltage of fixed phase can be obtained which is independent of frequency. As the oscillator is tuned over the range, the output of the first phase inverter is of constant phase but is displaced 90 deg from the input voltage. If these two voltages are applied to the horizontal- and vertical-deflection plates of a cathode-ray oscilloscope, a circular sweep is obtained which is maintained at all frequencies to which the oscillator is tuned. This is a very convenient arrangement for use in frequency measurements, particularly if the CRO has a radial deflection system. By adding a locking system, the oscillator can then be locked in fixed phase to standard frequencies, or multiples or submultiples thereof, giving great flexibility with highly accurate results.

A diagram of such an oscillator is given in Fig. 73.

43. Ultra-low-frequency Multiphase *RC* Oscillator. From the point of view of adjustment of frequency, range covered, control of amplitude, and purity of wave form, the three-phase *RC* oscillator has many points of interest.

The principle of the circuit as described by Smiley[5] is illustrated in Fig. 74.

One major factor in u-l-f oscillator design is the size and cost of the frequency-determining networks required to secure oscillations at frequencies so low that a single

[1] Frazer, W., Wide Range Phase Shift Oscillator, *Electronic Eng.*, March, 1956, p. 200.
[2] Rosenstreter, E. W., Single-signal Single-sideband Adapter, *Electronics*, July, 1948, p. 124.
[3] Villard, O. G., Tunable A-f Amplifier, *Electronics*, July, 1929, p. 77.
[4] Rao, P. Venkata, Phase Shifter to Test Power Meters, *Electronics*, January, 1956, p. 156.
[5] Smiley, Gilbert, Ultra-low-frequency, Three-phase Oscillator, *Proc. IRE*, April, 1954, p. 677.

cycle may extend over an interval of 100 sec. Interfacial polarization in mica dielectrics becomes serious at frequencies below 20 cps.

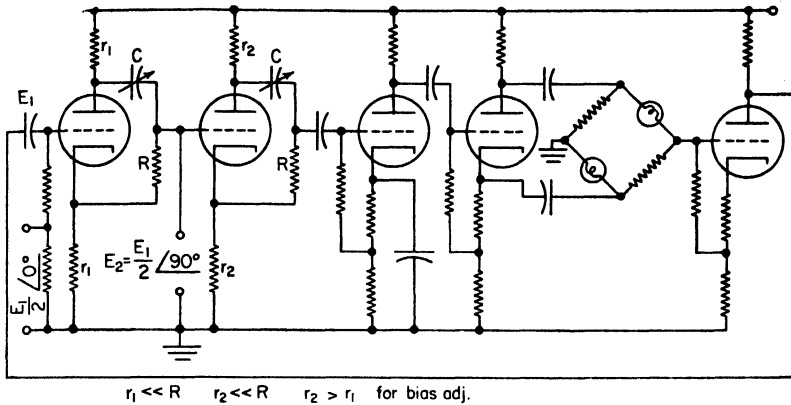



Fig. 73. Oscillator circuit giving two voltages 90 deg apart at any frequency to which oscillator is adjusted.

An ingenious solution is that of obtaining the required effectively large values of capacitance through the use of small capacitances, of satisfactory properties, multiplied by means of the "Miller effect."

In the very simplified circuit of Fig. 74, the three RC circuits are disposed in a three-phase Y connection around the power supply. Since the sum of the phase currents at the neutral of a balanced three-phase Y is zero, no phase currents of oscillator frequency traverse the power supply. The oscillator is therefore independent of the impedance of the power supply. The supply thus requires only a minimum of filtering, regulation, and by-passing.

The oscillator is basically a three-stage reentrant d-c amplifier. Oscillation is at that frequency for which the phase shift is 60 deg in each of the three identical RC networks. At this frequency, each network attenuates the signal by 2. The d-c loop gain must be at least $2^3 = 8$. At d.c. (zero frequency), there is no RC phase shift so the d-c loop gain gives an attenuation of 18 db, making the system exceptionally stable against d-c drift.

The plate circuit voltage dividers R_v are ganged to effect tuning by variation of the Miller-effect gain with a minimum gain of 2 per tube. The system shown would have a 3 to 1 tuning range and, for the lowest frequency, the capacitors need to be but one-third the value necessary if the Miller-effect multiplication were not used.

Fig. 74. Three-phase RC oscillator having several interesting features.

The details of the operating circuit are rather involved and will not be covered here. For all such information, reference should be made to the article by Smiley.

A possible use of such an oscillator is in the measurement of transmission and phase performance by using a network of resistive components only to obtain a voltage adjustable in magnitude and phase with reference to a fixed voltage.

Through use of a Scott connection utilizing resistors, a two-phase output can be obtained which will generate a true circular sweep on a CRO at any frequency to which the oscillator is adjusted.

RELAXATION OSCILLATORS AND MULTIVIBRATORS

44. Introduction. The pulses used to time the sequential action of circuits in many modern communication systems are generated, shaped, and controlled in time duration by *RC* timed relaxation oscillators. The most common forms of these oscillators are the multivibrators and the blocking oscillator. All these circuits have in common the features of poor inherent frequency stability, ease of triggering (synchronization to another source of frequency), high output voltage (or current), and high harmonic content in their output wave form.

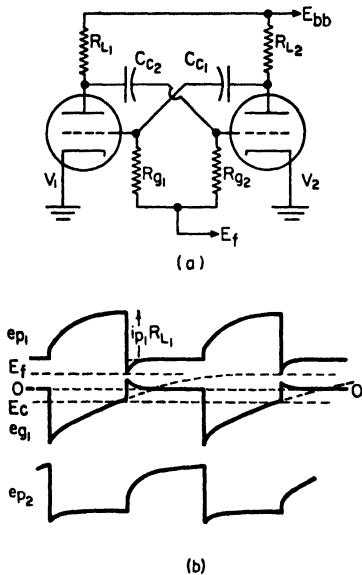



Fig. 75. Typical astable multivibrator, with wave forms in (*b*).

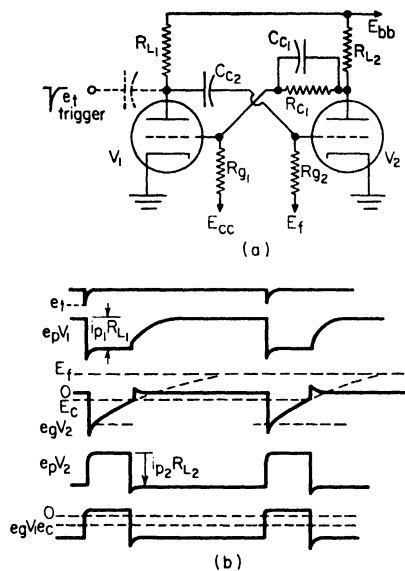



Fig. 76. Basic monostable circuit and wave forms.

Since these circuits are all similar in operation as far as stability, timing, and triggering are concerned, the next section will treat these common characteristics, and the succeeding section will discuss the individual characteristics of the more common forms of relaxation oscillators.

45. Multivibrator Basic Operation. The various forms of multivibrator circuits are most conveniently classified according to their time-stability characteristics. Usually there are two tubes involved in these circuits, and each tube has only two states of interest; it is either off or on. The off state is that in which the tube is biased beyond plate-current cutoff and in the on state the tube is commonly conducting near to zero-bias plate current. In the astable circuit, shown in typical form in Fig. 75, neither state of plate current can persist indefinitely. If tube V_1 is conducting, its plate current will drop a voltage in R_{L1} which is capacitively coupled via C_{c2} and R_{g2} to the grid of V_2. V_2 will be cut off until C_{c2} has discharged sufficiently to permit it to conduct. When V_2 begins to conduct, both tubes form a regenerative loop and V_1 is quickly switched off.[1] This state of affairs, with V_1 off and V_2 on, will persist

[1] Williams, E. M., D. F. Aldrich, and J. B. Woodford, Speed of Electronic Switching Circuits, *Proc. IRE*, **38**, 65, January, 1950. Chance, B., *et al.*, "Waveforms," M.I.T. Radiation Laboratory Series, Vol. 19, Chap. 5, p. 175, McGraw-Hill, 1949.

until C_{c1} discharges sufficiently to permit V_1 to conduct again. The timing sequence for this astable circuit is shown in Fig. 75*b*.

46. Monostable MV. Figure 76 shows a circuit revision which will convert the simple astable multivibrator to a *monostable* circuit. Without any external disturbance of the circuit, V_2 will be conducting near zero bias (grid returned to E_f through R_{g2} and cathode grounded). The drop in R_{L2} due to V_2 plate current is applied through R_{c1} and R_{g1} to keep V_1 biased beyond plate-current cutoff. This state is stable. Now upon the application of a triggering signal of appropriate polarity (a negative pulse is shown applied to the "on" grid in Fig. 76), the circuit will switch regeneratively. V_2 will go off and its rising plate voltage will drive the grid of V_1 up to the point where grid current will be drawn. The drop in plate voltage of V_1 is coupled via C_{c2} and R_{g2} to the grid of V_2 and will keep V_2 off for a time determined

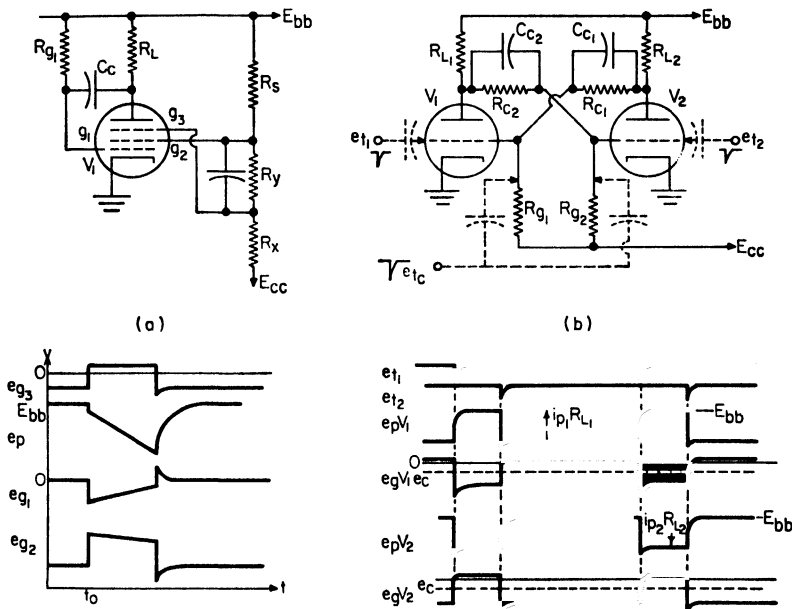



FIG. 77. (*a*) Simple phantastron circuit with wave forms. (*b*) Bistable circuit.

by the $R_{g2}C_{c2}$ time constant and V_1 plate swing. When C_{c1} discharges sufficiently to permit V_2 to conduct, the circuit will regeneratively switch back to its original stable state where it will remain until triggered. This timing and triggering sequence can be followed in Fig. 76*b*.

A second major class of multivibrators, usually having monostable characteristics, consists of those circuits which are timed by a Miller-effect sweep circuit.[1] The simplest form of this monostable circuit is the phantastron shown in Fig 77*a*.[2] This circuit form has many variations to fit it to particular applications. Chief among these are the cathode-coupled and cathode-follower-coupled phantastrons and the sanatron circuits.[3]

The most striking difference between the phantastron circuit of Fig. 77*a* and the astable and monostable circuits of Figs. 75 and 76 is in the means of timing. In the circuit of Fig. 77*a*, timing is based on the Miller effect.

[1] Chance *et al.*, *op. cit.*, pp. 195–204, 279.
[2] Millman, J., and H. Taub, "Pulse and Digital Circuits," p. 222, McGraw-Hill, 1956.
[3] Chance, *op. cit.*, p. 279.

The tube shown in Fig. 77*a* is one of those in which both grids 1 and 3 are sharp cutoff (6AS6, 5915, etc.). Normally, plate current is off because the high screen-grid current lowers the grid 3 below its base. Under these conditions, C_c will be charged to E_{bb} with G_1 drawing current through R_{g1}. Suppose the circuit is triggered by, say, a negative pulse on grid 1. Screen current is momentarily reduced, raising the screen voltage and hence the voltage at grid 3. This action is regenerative, for when plate current comes on, the drop across R_1 lowers grid 1. A negative step appears at the plate; and, since the voltage across C_c cannot change instantaneously, the same negative voltage step appears at G_1. G_1 obviously cannot cut off plate current; so an equilibrium is established at some low value of plate current. As C_c discharges, grid 1 tends to become more positive, increasing plate current and decreasing the plate voltage, replacing most of the lost charge on C_c. The equation for the plate voltage as a function of time depends upon R_gC_c and upon the gain from grid 1 to plate, K, the actual equation being[1]

$$e_p(t) = KE_{bb}\left[1 - \epsilon^{\frac{-t}{(1-K)R_gC_c}}\right]$$

which, for K large with respect to 1, reduces to the linear term of a series:

$$e_p(t) = -\frac{E_{bb}}{RC}t$$

Therefore, during the astable portion of the multivibrator cycle, the plate voltage runs down nearly linearly as grid 1 voltage increases. This action is terminated when the plate voltage "bottoms" and space current switches to the screen, lowering grid 3 voltage so that the plate current regeneratively switches off. After the astable portion of the period, C_c must recover its charge via the grid 1 cathode space and R_1. This exponential recharge is very similar to the recovery of the astable multivibrator of Fig. 75.

One obvious means of controlling the astable period of the phantastron is to control the plate voltage independently of the screen voltage. This is advantageous since the time to run down is linearly related to the starting-plate voltage.

The cathode-coupled version of the circuit eliminates the necessity for a negative supply in a fashion similar to the cathode-coupled monostable shown in Fig. 82. Recovery time can be improved by driving the coupling capacitor through a cathode follower. These various special circuit forms are analyzed in detail in Chance and in Millman and Taub.

47. Bistable MV. The bistable circuit constitutes the third major class of MV. A basic circuit is shown in Fig. 77*b*. As its name implies, the bistable circuit has two positions of stable equilibrium in which it can remain indefinitely. In order to be useful, it must be triggered. Let us first examine the stability conditions and then consider how the circuit can be triggered. V_1 is conducting and its low value of plate voltage can be caused to keep V_2 off by correctly choosing the R_{c2}, R_{g2} voltage divider and bias voltage. Now, if the circuit is symmetrical $R_{L2} = R_{L1}$, $R_{c1} = R_{c2}$, and $R_{g1} = R_{g2}$, and if V_2 were on, V_1 would be off. The circuit can be caused to trigger in either of two ways. Referring to Fig. 77*b*, triggering pulses can be applied separately to each grid, or in common to both grids. If trigger pulses are applied to both grids simultaneously, the different charges on C_{c1} and C_{c2} cause the circuit to change state, and under these circumstances, the circuit will divide the input frequency by 2.

48. Blocking Oscillator. A typical blocking-oscillator trigger circuit is shown in Fig. 78. By either grounding R_g or returning it to a bias supply adequate to cut the tube off, the circuit can be caused either to oscillate at the frequency determined by C_cR_g or to respond to an input positive triggering pulse.

To understand the basic principles of operation, assume that V_1 has been off for some reason and just begins to conduct. The rising plate current induces a voltage in the secondary of T_1 so as to drive the grid of V_1 positive, causing V_1 to conduct more

[1] Chance, *op. cit.*, p. 279.

heavily. This regenerative action quickly drives the grid highly positive. A typical situation finds the grid and plate at nearly the same potential at around half the plate-supply voltage. The heavy grid current charges C_c, and the voltage across C_c will go negative with respect to ground. The build-up of plate current in V_1 is limited by the primary inductance of T_1 and is at first nearly linear so that a nearly constant positive voltage is present at the grid. When C_c is charged to nearly the induced grid voltage, any decrease in plate current will cause the grid voltage to fall and the circuit will then switch off regeneratively. The charge on C_c must now be lost via R_g before conduction can begin again. The exact operation of the circuit is not nearly so simple as pictured here. Actually, there are several possible modes of operation for the circuit depending upon the transformer parameters and capacitance C_c.[1]

FIG. 78. Basic blocking-oscillator circuit and wave forms.

When the blocking-oscillator circuit is self-oscillatory, the free-running period is the sum of the "on" time and the time set by the exponential discharge of C_c through R_g (see following section). Since the "on" time is usually very brief (0.1 to 5 μsec), the period is very near to the time set by C_c, R_g.

RELAXATION OSCILLATOR CIRCUIT DESIGN

49. Basic Timing Equations. All the switching circuits described above have their

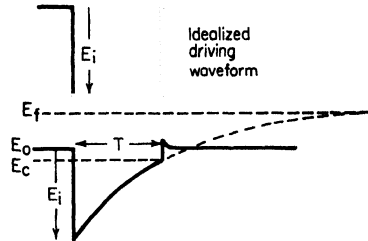



FIG. 79. Illustration of the timing principle.

astable periods timed in an identical way. That is, a tube goes into conduction and applies a step of voltage to a capacitor, which then must discharge to the critical point where the second tube involved (the same tube in the case of the blocking oscillator) can conduct. Figure 79 illustrates this timing operation. The driving wave form (magnitude E_i) is usually the plate swing of a tube. E_0, the initial value of voltage at which the timing grid is maintained, is usually zero and can be omitted. After the driving plate comes on, the grid voltage swings negative by the same amount as the plate. The equation for the grid voltage with time is given by[2]

$$e_g = E_f + (E_0 - E_i - E_f)\epsilon^{-\frac{t}{R_gC_c}} \tag{5}$$

Now at time T_p, the instantaneous grid voltage is $-E_c$ and

$$-E_c = E_f + (E_0 - E_i - E_f)\epsilon^{-\frac{T_p}{R_gC_c}} \tag{6}$$

[1] Benjamin, R., Blocking Oscillators, *J. IEE*, **93**, Part IIIA, No. 7, 1159, 1946. Millman and Taub, *op. cit.*, Chap. 9, pp. 253–285.

[2] Chance *et al.*, *op. cit.*, p. 178.

Solving this equation for the time,

$$T_p = R_g C_c \log_\epsilon \frac{E_f + E_i - E_0}{E_f + E_c} \tag{7}$$

This is the basic MV timing equation. It says that the period of the multivibrator due to the time constant is linearly proportional to $R_g C_c$ and proportional to the natural logarithm of the quantity given by the final value toward which the capacitor is discharging, plus the plate swing, divided by the final value of grid voltage plus the critical grid voltage at which the tube will again switch state. When the final value is ground and when the initial value is ground, this equation reduces to

$$T_p = R_g C_c \log_\epsilon \frac{E_i}{E_c} \tag{8}$$

and the logarithmic multiplier is just the driving swing divided by the cutoff voltage.

In preparation for the practical design of multivibrators, several important facts can be learned from this equation. First, there is a first-order independence of the period with changes in the plate supply for the circuit. This is because both the plate swing of the driving tube and the cutoff of the driven tube are directly proportional to plate-supply voltage: $E_c = E_{bb}/\mu_{co}$, and $E_i = E_{bb}\bar{r}_p/(\bar{r}_p + R_L)$, where $\bar{r}_p$ is the static plate resistance of the driving tube. Using these relationships, Eq. (8) becomes related to the tube and load resistor.

$$T_p = R_g C_c \log_\epsilon \frac{\bar{r}_p \mu_{co}}{\bar{r}_p + R_L} \tag{9}$$

The stability of timing is therefore proportional to variations in these quantities, of which $\bar{r}_p$ for the tube is usually the largest variable during the life of the tube. Methods of stabilization of the period against this variable are given in Sec. 57.

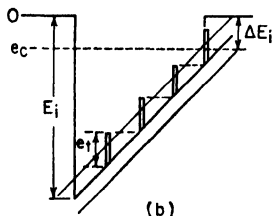
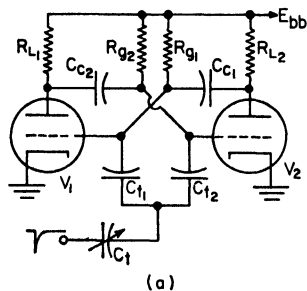



FIG. 80. (*a*) Practical astable multivibrator connected for frequency division. (*b*) Idealized timing wave form showing principles of frequency division.

50. Frequency Division Using Synchronized Multivibrators. By adding pulses at a higher frequency to the grid-timing wave form, the astable or monostable multivibrators can be locked to some submultiple of the higher frequency. Figure 80*a* shows an astable multivibrator circuit similar to Fig. 75 connected[1] as a frequency divider. Figure 80*b* is a linearized drawing of its grid-timing wave form showing the principles of division. Triggering pulses of amplitude e_t with period T_t are added to the timing wave form. (The usual procedure is to add negative triggering pulses to the "on" grid of the multivibrator as shown in Fig. 80*a*.) In the example shown, the division ratio is 4. In designing the MV to divide most reliably the quantities which must be exactly determined are the trigger amplitude as determined by the magnitude of the plate swing E_i and the division ratio R. Examination of Fig. 80*b* will permit one to write

$$E_t = \frac{2(E_i - E_c)}{2N + 1} \tag{10}$$

A simple logical argument will show that this is the correct trigger amplitude for

[1] Shenk, E. R., Multivibrator, Applied Theory and Design, *Electronics*, **17** (1), 136; (2), 140; (3), 138, 1944. Millman and Taub, *op. cit.*, Chap. 12, pp. 355–372.

maximum stability. The results of this derivation are equally applicable in the actual case where the timing wave form is an exponential.

The two quantities which tend to vary over the life of the frequency-dividing MV are E_i, the magnitude of plate swing, and the slope of the timing wave form set by R_g and C_c. Suppose the driving tube ages; E_i will decrease by just $\frac{1}{2}e_t$, at which time the peak of the $N - 1$ pulse will reach E_c and the circuit will fail. If e_t were *larger* in amplitude, E_i could not decrease as much before failure. Now suppose that for some reason E_i increases. It can increase by exactly $\frac{1}{2}e_t$ before the division fails by jumping to $N + 1$. If e_t were *less* in amplitude than Eq. (6) dictates, the unit will fail before E_i increased by $\frac{1}{2}e_t$ by the peak of e_t not passing through E_c. This same argument holds true for changes in R_gC_c, causing the timing slope to change with E_i fixed at its correct value.

Now that the optimum triggering conditions have been derived, it is apparent just how much variation in E_i and slope can be tolerated before a divider will fail. In the linearized example of Fig. 80*b*, with $N = 4$, suppose $E_i - E_c$ is 100 volts; then e_t should be 22.2 volts. The division will fail when E_i decreases by about 11 volts. Thus, about ± 10 per cent can be tolerated in the plate swing of the driving tube with a 4:1 division. The fact that the timing wave form is truly exponential *decreases* this tolerance because the slope near the critical voltage is *lessened*. Apparently, to build multivibrators to divide safely by ratios of even 4:1 or 5:1, precautions must be taken to stabilize the driving plate current and to determine slope correctly either by using close-tolerance components or by exactly adjusting the R_gC_c product. Since even new tubes vary by as much as 20 per cent in plate current and μ_{co}, to divide by 5:1 on each side of an astable MV requires adjustment of both RC products and the trigger amplitude. Again the methods of stabilizing plate current in Secs. 53 to 57 should be applied in the design of reliable frequency-dividing multivibrators.

51. Harmonic Generation. The abrupt changes in voltage, occurring in all forms of the trigger circuits discussed here, when switching from one state to another make them useful as sources of harmonics for frequency multiplication, for heterodyne calibration, etc.[1] A great deal must be known about the circuit and its wave forms before the exact harmonic spectrum can be predicted by Fourier analysis.[2] In general, if the maximum extent of harmonic spectrum is desired, the circuit must be designed to produce one voltage transition in the minimum possible time. For example, design of the switching circuit to produce square waves will yield an output spectrum with a minimum of even-order harmonics.

TRIGGER CIRCUIT DESIGN

52. Basic Requirements. Any one of several qualities of the trigger circuit can predominate to dictate the practical design procedure. The circuits must all be designed to meet the basic requirements for correct operation as outlined below, but these basic requirements must be modified to fit the circuit to the system in which it is a component. Practical requirements which usually dictate circuit design individually or in combination are:

1. Wave shape (rise or fall time, flatness, etc.)
2. Speed (of either transitions or maximum repetition rate)
3. Output voltage
4. Output impedance

Together with these basic requirements is that of reliability. With the increasing complexity of today's electronic systems, this requirement often becomes predominant, limiting performance in all other categories. Designing reliability into the trigger circuit goes much further than just choosing the best components and most reliable tube types. Wherever the circuit can be made to meet its other basic requirements,

[1] Frank, R. W., A High-precision Calibrator for Frequency and Time, *Gen. Radio Experimenter*, **31**, 4, June, 1956.

[2] Kandoian, A. G., *et al.*, "Reference Data for Radio Engineers," pp. 1019*ff.*, Federal Telephone and Radio Corp., 1956.

the residual performance can all be used up in obtaining reliability. When a particular circuit cannot meet its performance requirements with the desired margins for tube aging, or component drifts or tolerance, a better tube, more costly components, or even another stage must be added. Finally, when the best available tubes have been applied, reliability must limit maximum attainable performance or vice versa.

53. Design of the Astable Circuit. The astable MV is commonly used as a simple and economical oscillator, or as a frequency divider. The maximum frequency and the stability of the output frequency (or in the case of the frequency-dividing MV the reliability of the division as the tube ages) are the most important criteria.

A straightforward qualitative analysis of the circuit dictates the design procedure. Refer again to Fig. 75. The plate current of the "on" tube produces a drop in the load resistor R_{L1} which keeps the "off" tube off for a time determined by Eq. (7). After the second tube switches on, turning the first off, the coupling capacitor and stray capacitances must recharge via R_{L1} and the grid resistance of the "on" tube. This recharge cycle will limit the maximum frequency for a given C_c. To obtain maximum frequency, C_c must be lowered as far as practical, a limit being imposed by the input capacitance at the grids which will reduce the timing voltage. Having lowered the C_c, one must then decrease R_L and use a higher-perveance tube to maintain an adequate plate swing. A practical limit to this procedure is obtained when the best tube has been used. A typical circuit in the form of Fig. 80 for operation at around 1 Mc uses a type 5687 tube with 5,100-ohm plate resistors, C_c = 33 $\mu\mu$f, and R_g = 43,000 ohms. With a 150-volt power supply, a plate swing of 95 volts will be obtained. In this limiting example, the frequency stability of the circuit over the life of the tube is poor. Figure 81 shows what happens to the plate swing as the tube ages and $\bar{r}_p$ doubles at end of life. The end-of-life plate swing is 70 volts, and from Eq. (7) the frequency will rise by about 22 per cent. A means for stabilizing this circuit against tube aging is given in Sec. 57.

FIG. 81. Characteristics of type 5687 tube showing effect of aging.

At lower frequencies, much more stable operation is possible.[1] Consider the design of a symmetrical multivibrator for 10 kc. R_L is chosen at 100,000 ohms; this sets C_c at 125 $\mu\mu$f if four time constants of recovery of plate voltage for the off tube are allowed. R_g must then be 680,000 ohms. Even if the tube chosen is a 12AU7, the plate swing for a new tube is 140 volts, for an end-of-life tube 130 volts, and the "unlocked" frequency will only vary about 4.5 per cent down as $\bar{r}_p$ doubles. With this sort of stability, the circuit could be used to divide frequency reliably up to 20:1, if no other variables but $\bar{r}_p$ for the tube are considered!

54. Design of Monostable Circuits. The timing portion of the cycle of the monostable circuit obeys the same rules as the astable circuits just discussed. Both the monostable and bistable circuits must fulfill certain d-c conditions for proper operation over the life of the tubes. The grid of V_1 in the monostable circuit of Fig. 76 under all conditions of component tolerance, supply variation, and tube aging must be maintained beyond cutoff. This can be assured by choosing an R_{L2} appropriate

[1] Harris, C. C., The "Hard-bottoming" Technique in Nuclear Instrumentation Circuit Design, *IRE Trans. Nuclear Science*, NS-3 (2), 5, March, 1956.

to the desired output voltage, impedance, or rise time (thus establishing, by this combination, the tube necessary to meet both output-voltage and band-width requirements). Now having V_2 and R_{L2}, the tube can be aged in imagination to the desired "end-of-life" conditions, say to $\bar{r}_p$ doubled. This end-of-life plate swing and the worst combination of component tolerances must establish cutoff in V_1. At the same time, when V_2 is off, V_1 must be brought up to the desired level of plate current (usually zero bias). This swing places a further requirement on R_{L2} or the negative supply. R_{L1}, R_{g2}, and C_c can be chosen on the same basis of recovery time and astable timing as for the astable circuit.

FIG. 82. Circuit and wave forms of cathode-coupled monostable multivibrator.

The circuit of Fig. 82 shows a somewhat more sophisticated monostable. This cathode-coupled circuit is the most commonly encountered form in practice.[1] It both eliminates the negative supply and stabilizes the timing plate swing of V_1 so long as V_1 is not permitted to swing into grid current. In its stable quiescent condition, V_2 is on and in grid current. The zero-bias plate current of V_2 holds the cathode potential at a sufficiently high level that V_1 is off. This must be ensured even at "end of life" for V_2. When the circuit is switched, say by application of a negative pulse at the plate of V_1, V_2 goes off and V_1 comes on as the cathode voltage falls. As long as V_1 is not in grid current, the effective resistance seen looking back into its plate is

$$R_{\text{eff}} = \bar{r}_p + R_k(1 + \mu) \tag{11}$$

With freedom to choose R_k and with the choice of a high-μ tube, $R_k(1 + \mu)$ can be made very large with respect to $\bar{r}_p$ and variations in $\bar{r}_p$ will cause very little variation in plate current.

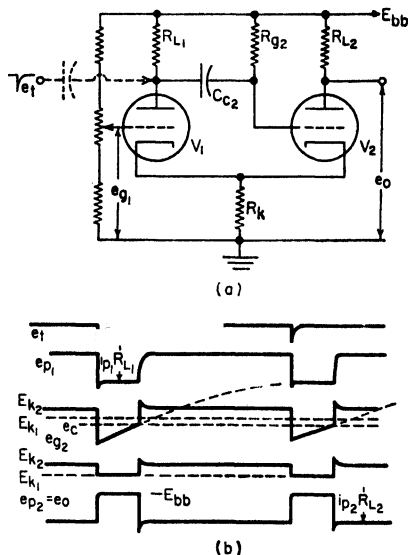
This circuit has another valuable characteristic. Because of the linear relationship of the control grid voltage and the plate current of V_1, the timing of the astable period is highly linear with respect to the control-grid voltage.

In the design of this form of monostable MV, R_{L2} is chosen to provide the desired output voltage. $R_{L2} + R_k$ with V_2 in grid current established by R_g are chosen to set the operating current for V_2 within its ratings. As usual the choice of impedance level and the tube used will determine the switching speed. $R_{L1} + R_k$ with the bias on V_1 when it is on establish the plate swing of V_1. This is the timing swing E_i of Eq. (8). It is now only necessary to choose C_c and R_{g2} to establish the astable period.

After the astable period terminates, C_c must reestablish its charge through the grid-cathode resistance of V_2, R_k, and R_{L1}. R_k, R_{L1}, and C_c should be set so that at least four time constants,

$$4(\bar{r}_g + R_{L1} + R_k)C_c \tag{12}$$

are allowed before retriggering at the highest repetition rate desired; otherwise there will be a large effect on the timing.

55. Bistable-circuit Design. Perhaps no other circuit in use in present-day electronics has had so much effort spent in attaining optimum design as the bistable multi-

[1] Chance *et al.*, *op. cit.*, p. 170. Millman and Taub, *op. cit.*, pp. 187*ff*.

vibrator.[1] By consulting the literature, the circuit designer can obtain a design optimized for the particular plate swing, speed, or life desired.

As usual, the speed is determined by the quality (g_m) of the tube chosen. A 12AU7 can deliver 100-volt plate swings at speeds up to 100 kc quite reliably, a 5687 up to 1 Mc, and beyond that, high-quality pentodes can be operated up to speeds in excess of 10 Mc with small plate swings. Special techniques have produced bistable circuits capable of operation at speeds in excess of 50 Mc.[2]

Where a negative supply is available, a bistable circuit of simplest form (Fig. 77) can be designed. Having chosen the tube and its load resistor for the desired plate swing, the R_cs and R_gs are chosen to ensure that (1) when the driving tube is off, the driven tube will be in grid current, and (2) when the driving tube is on, the driven tube will be driven beyond plate-current cutoff. Both conditions (1) and (2) must be met under the worst possible circuit conditions of resistance tolerance, expected tube aging to "end of life," initial variations in the cutoff μ for the tubes, and supply variations. All these variations require that the *average* tube be highly overdriven and this will reduce the switching speed. Therefore, in attaining maximum speed, it is customary for as many of these variables as possible to be eliminated or reduced. Often it is necessary and economical to use precision resistors in the cross-coupling networks, regulated supplies, and specially built (computer) tubes whose μ_{co} and initial-plate-current balance are carefully controlled.

The most common form of the bistable eliminates the negative supply by using a common cathode resistor. The drop across this resistor is the value of the negative supply and the design can proceed as described in the last section. $R_k + R_L$ is first laid out as a load line on the tube characteristics and the desired plate current and output voltage are established. The values of the cross-coupling dividers are then set to ensure zero bias and cutoff d-c conditions under the worst expected set of circuit conditions.

56. Triggering of Bistable Circuits. As stated in the introductory paragraphs, there are two distinctly different ways in which it is desirable to trigger the bistable circuit. It is often desired only to set the circuit into one state or the other. In Fig. 77, a negative pulse T_1 will set V_1 off and V_2 on, while negative pulse T_2 will reverse the state. This type of "single-ended" triggering requires very little a-c coupling via C_{c1} and C_{c2} from plate to grid and these capacitors can be set at a value to just compensate for the input capacitance of the tubes. The second sort of trigger injection is that in which the trigger pulse is applied to both grids simultaneously. The circuit is said to be "complemented." It is this form of triggering which must be used when the circuit is to divide the input trigger rate by 2. In this case, the stored charge on C_{c1} and C_{c2} causes the circuit to change from one state to the other with each trigger. The capacitors must be increased to give an effective *overshoot* at each grid lasting for a longer time than the trigger pulse. Therefore, after each triggering pulse, the grid transient puts a larger voltage on the grid than that predicted by the d-c design. It is the recovery of this transient which will now limit the maximum speed at which the unit can be driven with a given trigger pulse. The cross-coupling capacitors must thus be chosen to permit reliable triggering under the worst trigger-pulse amplitude conditions, while still not limiting the maximum speed below that desired. The maximum speed due to the recovery of this transient can be raised by reducing the time constant R_cC_c at the expense of increased power dissipated in the cross-coupling network and a lowered effective value of E_{bb}.

57. Stabilized-plate-current Multivibrators. In all the multivibrators discussed in the preceding paragraphs, the demands for increased speed call for a lowering of circuit impedance levels and tubes of higher perveance to provide the desired plate swings. With impedance levels approaching the static plate resistance, the plate swing and therefore the timing or d-c stability become highly dependent on plate resistance, and thus upon tube aging. In Sec. 54 it was pointed out that the timing due to the plate swing of the left-hand tube of Fig. 82 was nearly independent of the

[1] Millman and Taub, *op. cit.*, Chap. 5, pp. 140–173. Ritchie, D. K., The Optimum DC Design of Flip-flops, *Proc. IRE*, **41** (10), 1614, November, 1953. Pressman, R., How to Design Bistable Multivibrators, *Electronics*, April, 1953, p. 164. Paivinen, J. O., and I. L. Averbach, Design of Triode Flip-flops for Long-term Stability, *IRE Trans. PGEC*, **EC-2** (2), 14, June, 1953.

[2] Bay, Z., and N. T. Grizamore, High-speed Flip-flops for the Millimicrosecond Region, *IRE Trans. PGEC*, **EC-5** (3), 121, September, 1956.

aging characteristics of the tube so long as the tube *did not* reach zero bias, the change of plate current with $\bar{r}_p$ being given by Eq. (11).

This principle can be extended to all the circuits of this class. A current-stabilized astable circuit is shown in Fig. 83. The grids are prevented from reaching zero bias during the expected tube life by the diode grid clamps. Crystal diodes can be used so long as the grid impedance levels are sufficiently low that back resistance can be neglected at all temperatures. The desired timing stability can be designed into the circuit by choosing the R_k appropriate to Eq. (11).

FIG. 83. Current-stabilized astable multivibrator.

The extension of this principle to the bistable circuit and other monostable forms is obvious. In all forms of the stabilized circuits, the concomitant advantage of a stabilized amplitude for the output wave form is gained.

Any other means of stabilizing the plate swing of the timing tube will, of course, stabilize the circuit against tube variations. Often because of the undesirability of adding capacitance at a grid or the desirability of using semiconductor diodes, it is possible to stabilize a plate swing by d-c clamping techniques. The disadvantage of this system lies in the low impedance required of the clamp supply voltage and the necessary additional power. If many similar circuits are used in the system, a separate power supply for the clamps can be an economical solution.

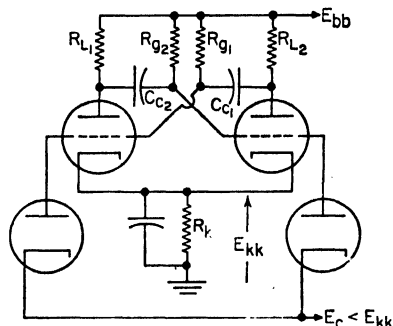
58. Blocking-oscillator Design. The design of a satisfactory and reliable blocking-oscillator circuit hinges primarily on the transformer used. The design of pulse transformers is beyond the scope of this section, and the reader is referred to the literature.[1] Many manufacturers of pulse transformers publish a recommended circuit and offer a line of pulse transformers to produce a wide gamut of possible pulse durations when these transformers are used in their circuit. As a general rule, the most desirable tubes to be used in a blocking oscillator have a low ratio of control-grid current to plate current when the control grid is highly positive. These are usually low-mu tubes, such as the 12AU7, 5687, 6350, etc.

There always arises the question of how the tube will stand up under the heavy instantaneous pulses of current encountered in blocking-oscillator service. Very few life-test data are available for this class of service. Experience indicates that, so long as the pulses are sufficiently brief that space charge is not exhausted, only the usual ratings of average cathode current, element dissipations, and peak voltages need to be met. For the most conservative designs, tubes such as the 6350 and 6463, which are assigned peak current ratings for all elements, can be used with confidence.

TRANSISTOR OSCILLATORS

Transistor oscillators may be divided into two different classes, negative-resistance oscillators and feedback oscillators. Oscillators of the first type make use of the negative driving-point impedance exhibited by point-contact transistors with certain circuit arrangements. While oscillators of this type have been considered at some length,[2] only the mechanism of obtaining the negative resistance is different from previously known oscillators of this type, and as junction-transistor characteristics are steadily improved, negative-resistance transistor oscillators are dwindling in importance.

[1] Glasoe, G. N., and J. V. Lebacqz, "Pulse Generators," M.I.T. Radiation Laboratory Series, Vol. 5, pp. 575*ff*., McGraw-Hill, 1948. Benjamin, *loc. cit.* Millman and Taub, *loc. cit.*

[2] Farley, B. G., Dynamics of Transistor Negative-resistance Circuits, *Proc. IRE*, **40**, 1497, November, 1952. Lo, A. W., R. W. Endres, J. Zawels, F. D. Waldhauer, and C-C. Cheng, "Transistor Electronics," pp. 353–373, Prentice-Hall, 1955.

59. Transistor Feedback Oscillators. In these circuits some of the output of an amplifier is fed back to its input in the proper phase to maintain oscillations. Transistor feedback oscillators are in many ways analogous to vacuum-tube feedback oscillators, but there are a few important differences:

1. The transistor generally has a low input impedance and behaves as a current amplifier, so that it is usually necessary to feed back a *current* of the proper phase rather than a *voltage*. As a very simple example, consider the phase-shift oscillator of Fig. 84. In the vacuum-tube case, the voltage across the last shunt resistor in the phase-shift network is applied to the grid of the first tube as in Fig. 84*a*, while in the transistor oscillator, the current in the last shunt resistor flows into the base of the first transistor as in Fig. 84*b*. When feedback is obtained from a tap on a tuned circuit, the low and generally nonlinear input impedance of the transistor must be considered.

FIG. 84. Phase-shift oscillators. (*a*) Tube type; (*b*) circuit using transistors.

2. Transistor dynamic characteristics are very linear up to the point of abrupt discontinuity (reverse collector bias or emitter-current cutoff) so that some external amplitude-limiting arrangements are necessary in almost all cases for low-distortion sine-wave output. Such power-sensitive devices as thermistors or lamp bulbs[1] have been included in oscillator circuits in such a way as to effectively decrease the positive feedback as the oscillation amplitude increases. Very effective amplitude control is obtained when such a control element is included in the Meacham bridge[2] circuit. Amplitude control of class C oscillators has been obtained by means of a bias voltage proportional to the output amplitude which is fed back to decrease the conduction angle as the output amplitude increases.[3]

3. Phase shift occurs in the current gain of a transistor at much lower frequencies than in the transconductance of a vacuum tube. Therefore, in transistor r-f oscillators the phase shift required of the feedback network may be any value from 0 to 360 deg. This effect also acts to decrease the span of the upper ranges of variable-frequency oscillators.

Transistor counterparts of nearly all types of vacuum-tube oscillators have been constructed. Crystal-controlled transistor oscillators have been built for frequency-standard purposes[2,4] and in this use transistors have the advantage, because of their

[1] Sulzer, P. G., Low-distortion Transistor Audio Oscillator, *Electronics*, **26**, 171, September, 1953.
[2] Witt, S. N., Jr., Transistorizing Meacham-bridge Oscillators, *Electronics*, **29**, 193, March, 1956.
[3] Kretzmer, E. R., An Amplitude Stabilized Transistor Oscillator, *Proc. IRE*, **42**, 391, February, 1954.
[4] Sulzer, P. G., Transistor Frequency Standard, *Electronics*, **26**, 206, May, 1953.

small size and low power requirements, that the entire oscillator circuit can easily be placed in a constant-temperature environment. A transistorized f-m signal generator has been constructed for the frequency range 20 to 100 Mc.[1] It makes use of a transistor oscillator operating at frequencies up to 33.3 Mc and a harmonic selector (frequency multiplier). Modulation is accomplished by a reactance transistor. An interesting application of transistor oscillators has been in the field of d-c to a-c conversion.[2] Such circuits make use of a push-pull square-wave oscillator wherein high over-all efficiency and very high collector efficiencies are obtained. Multivibrators[3] and blocking oscillators[4] have also been constructed with transistors. This circuit of a simple free-running MV is shown in Fig. 85. In this case there is a very close analogy to the corresponding tube circuits.

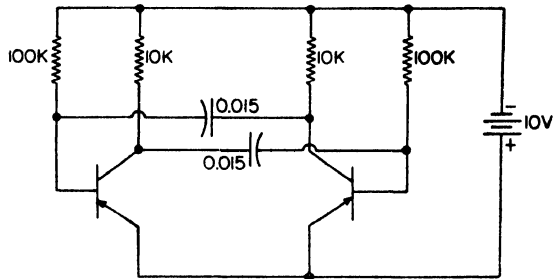



Fig. 85. Free-running transistor multivibrator using *p-n-p* junction transistors.

LOCKED OSCILLATORS

60. Frequency Dividers. Any of the *RC* oscillators previously described is capable of being locked to a control frequency signal, either of the same frequency or of a higher frequency. This effect is sometimes most troublesome if the filtering of the power supply is not adequate. The *RC* oscillator may then show definite locking ranges at multiples or submultiples of the line frequency.

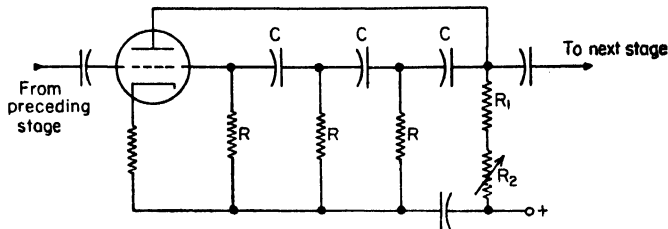



Fig. 86. Phase-shift oscillator used as a frequency divider.

In Fig. 86 is shown an arrangement for division from a 1,800-cycle tuning fork to 60 cycles. Two stages are used, one dividing by 5 and one by 6. The circuits are identical; only the values of *C* and *R* are altered.

61. Locking Circuits. More rigid control could be obtained by use of a tube as a voltage-controlled resistance element whereby the oscillator frequency can be changed by changing the grid voltage. If the grid voltage is derived from the output of a phase detector, then rigid control of the oscillator frequency can be obtained.

Since properly operating oscillators have no great harmonic content, it is necessary to generate harmonics by use of external circuits if reliable control of the oscillator is required when the oscillator fundamental frequency is a submultiple of the control frequency.

[1] Hupert, J. J., and T. Szubski, Transistorized F-m Signal Generator, *Electronics*, **28**, 133, February, 1955.
[2] Uchrin, G. C., and W. O. Taylor, A New Self-excited Square-wave Transistor Power Oscillator, *Proc. IRE*, **43**, 99, January, 1955.
[3] Prugh, T. A., Junction Transistor Switching Circuits, *Electronics*, **28**, 168, January, 1955.
[4] Linvill, J. G., and R. H. Mattson, Junction Transistor Blocking Oscillators, *Proc. IRE*, **43**, 1632, November, 1955.

A simple means, which is many times sufficient, is shown in Fig. 87. The voltage-controlled resistor could be used to replace R_2 of Fig. 86, for example.

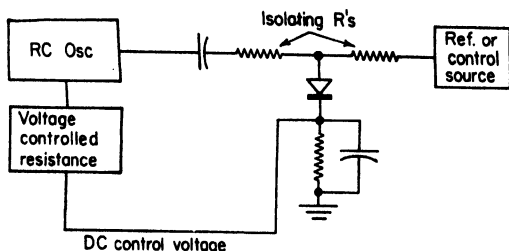



FIG. 87. An *RC* oscillator phase-locked to a reference frequency source.

Such a control system is very useful in obtaining submultiples of a standard frequency down to values in the range of 1 to 10 cycles.

For certain purposes, it would be desirable to lock a heterodyne frequency meter or signal generator to some standard frequency to avoid the necessity of continually checking the frequency to avoid errors of measurement due to drift.

Where oscillators draw grid current, a d-c control voltage can easily be applied through a series resistor R as indicated in Fig. 88. If R is of the order of R_g, an appreciable part of the d-c control voltage can be applied to the oscillator grid. Change of the d-c bias affects the frequency slightly, sufficient to give locking over a range generally more than adequate to maintain the oscillator in control for all ordinary temperature or supply-voltage variations over periods of an hour or more.

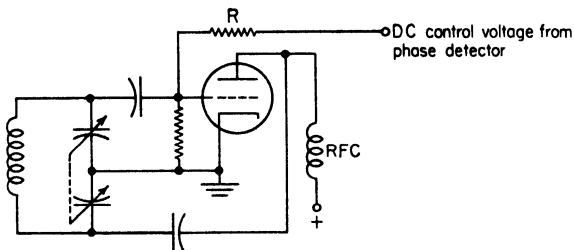



FIG. 88. Phase-locking a frequency meter to a reference frequency source.

BEAT-FREQUENCY OSCILLATORS

Beat-frequency oscillators are oscillator circuits in which the output signal is produced by beating together two signals differing in frequency by the output frequency. The two beating signals can be obtained from *LC* tuned oscillators, *RC* oscillators, or combinations of various types of oscillators. Among the various types of oscillators which are often used in beat-frequency oscillators are (1) two *LC* tuned oscillators, (2) one *LC* tuned and one crystal-controlled oscillator, (3) two klystrons. Other combinations, such as two *RC* oscillators or one *RC* oscillator and one crystal-controlled oscillator, sometimes occur, and at least one oscillator used three *LC* tuned oscillators to produce two beat notes simultaneously. The basic elements of a BFO are shown in the diagram of Fig. 89.

FIG. 89. Basic beat-frequency oscillator.

62. BFO Characteristics. Two distinct virtues not obtainable in other oscillator circuits are possessed by BFOs: (1) the output frequency can be made to cover many octaves of frequency on one dial range, and (2) a calibrated incremental frequency adjustment can be provided at any frequency by adding a small calibrated tuning adjustment to the fixed oscillator. The first feature is extremely valuable in work requiring recording of characteristics throughout the audio, or even the entire video, portion of the spectrum at one sweep; the second provides means for investigat-

ing narrow-band phenomena such as filter characteristics, which may occur at frequencies distributed throughout the spectrum covered by the oscillator.

Many other special tuning characteristics, such as multidecade logarithmic and linear-frequency scales including zero, are obtainable.

A constant output level can be obtained either by applying a-g-c techniques or by making the fixed oscillator operate to inject a smaller signal into the mixer than that from the variable oscillator. In the latter case, the beat note is controlled in amplitude by the smaller signal, which is fixed, and hence a constant-amplitude beat note results. Equally constant output results if both fixed and variable oscillators give constant input to the mixer at all settings of the dial.

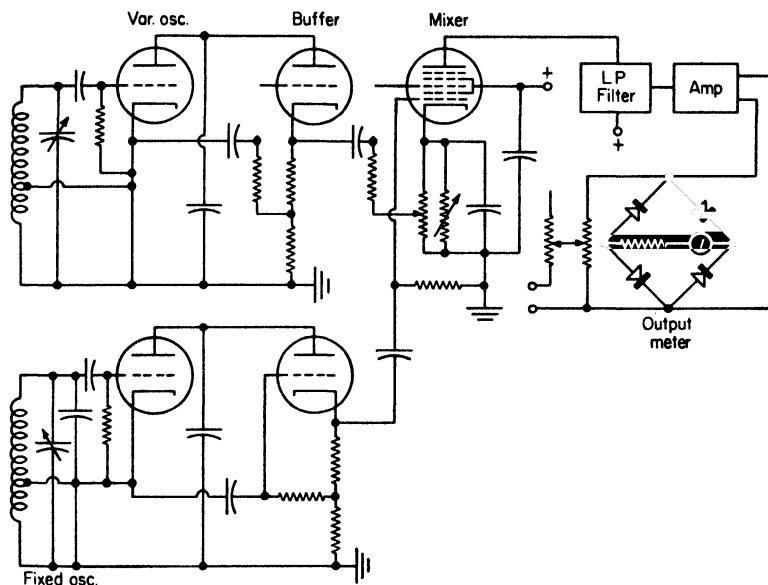



FIG. 90. Typical beat-frequency oscillator.

Harmonic distortion in the beat note results mainly from two causes, (1) distortion in the beating-oscillator wave forms, and (2) distortion in the mixing process. Distortion in the beating-oscillator wave forms becomes especially large if the oscillators are not adequately isolated by shielding and power-supply filtering when the oscillators are operating at frequencies giving low-frequency beat notes. Oscillators with large amounts of harmonic distortion in their wave forms tend to synchronize or lock under such conditions.[1] The a-f beat note produced by such unstable oscillators usually contains much harmonic distortion. The remedy is to reduce the harmonics in the oscillators and to isolate each from the other electrically. Distortion in the mixing process may be reduced by employment of a low-distortion mixing stage or by resort to balanced modulators to suppress unwanted higher-order mixing distortion products. The mixer stage shown in Fig. 90, using a 6SA7 tube, may be adjusted to produce a sine wave of total distortion under 0.08 per cent. Such a result requires ability to adjust the magnitudes of each of the beating voltages and of the bias values on each of the input grids. However, the purity of wave form easily obtainable by this means is equalled only by elaborate feedback systems in *RC* oscillators.

[1] Rouche, N., Étude du régime de synchronisation de deux oscillators couplés, *Bull. sci. AIM*, January, 1956, No. 1.

Oscillator beat notes other than the desired beat note, commonly called "birdies," are generally the result of harmonic distortion in one or both of the beating oscillators but are sometimes the result of distortion in the mixer or of undesirable pass bands in the filter or amplifier combined with small amounts of distortion in the desired beat note.

63. Stability. Since both oscillators may be made similar in construction, it is possible to minimize drift with temperature change by obtaining nearly identical frequency-vs.-temperature-time curves for both oscillators. If each oscillator is constructed to be very stable initially, the stability of the beat frequency may be extremely good. The interpolation oscillator manufactured by the General Radio Company (Fig. 91) covers the output-frequency range of 0 to 5,000 cycles with linear dial over

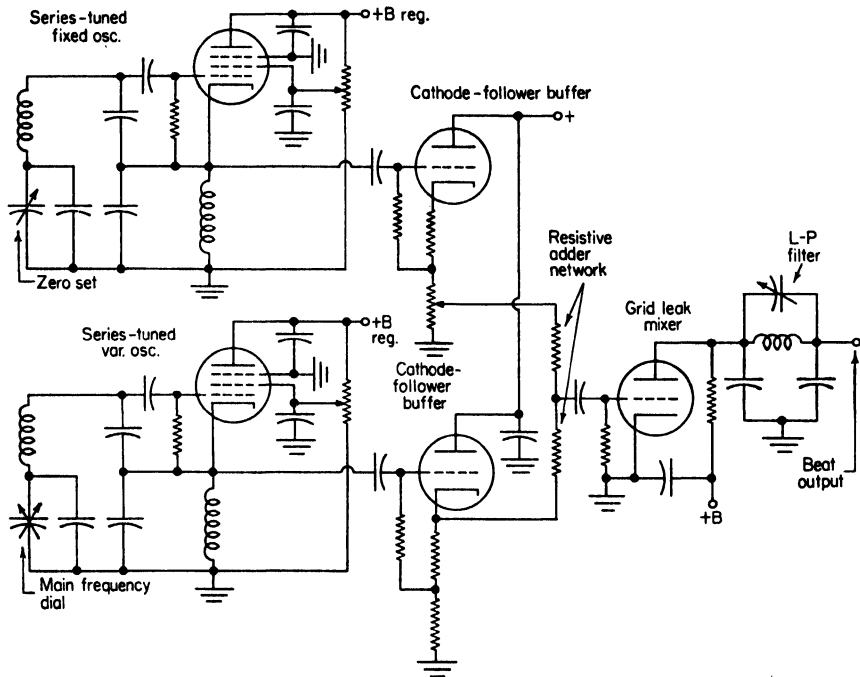



FIG. 91. Simplified schematic of frequency-stable beat oscillator.

the entire range. The maximum drift of the oscillator frequencies in this oscillator is less than 5 cycles from a cold start, and the output frequency will stay within ±0.5 cycle of its set value for hours of operation at room temperature. No temperature control is used. Both fixed and variable oscillators use the series-tuned circuit for maximum stability, the variable oscillator tuning from 42 to 47 kc, and the fixed oscillator being set at 42 or 47 kc by means of a panel switch.

The output-frequency stability of wide-range audio oscillators of the beat-frequency type is hardly less impressive, since it is easy to add negative-temperature-coefficient capacitor compensation to the r-f oscillators. A commercial BFO covering the nominal range from 20 to 20,000 cps shows less than 7 cycles drift of the output frequency in the first hour from a cold start, and generally stays within ±1 cycle of a given setting once it has operated for 1 hr. Compared with the *RC* oscillator which exhibits a stability proportional to frequency, this beat-frequency-oscillator stability is relatively poorer only below 100 cps, approximately equal up to 1,000 cps, and far superior

above 5,000 cps. Furthermore, the provision of zero-setting and cycles-increment adjustments on the BFO enable small frequency changes to be corrected without touching the main dial. Such an adjustment is not readily added to other oscillators.

Beat-frequency oscillators to cover the entire video-frequency spectrum have been constructed, the limiting factor usually being the availability of wide-band low-distortion amplification to follow the mixer. Tuned, tracked amplifier systems may be employed, but their difficulty of adjustment, and the lack of extremely wide-band tuning networks, usually results in the selection of the simpler wide-band amplifiers. The application of distributed amplifiers to this problem may provide even wider-range beat-frequency oscillators than have been employed heretofore.

The range of oscillator frequencies used is very wide. Special low-frequency beat oscillators for frequencies below 100 cps[1] have been made using oscillators below 5,000 cps. Commercially available beat-frequency oscillators using frequencies of 42 to 47 kc, 190 to 230 kc, and 39 to 49 Mc are now in production, and klystron BFOs beating two X-band (10,000-Mc) klystrons are also available commercially. As is apparent from this incomplete listing, the beat-frequency oscillator is a device of great versatility which enjoys wide application.

[1] Fleming, Lawrence, and William W. Follin, Subaudio Oscillator Tunes 0 to 50 Cycles, *Electronics*, **28**, 144, October, 1955.

CHAPTER 17

POWER-SUPPLY SYSTEMS

By Reuben Lee[1]

1. A-c vs. D-c Power Supplies. Vacuum tubes and transistors used in radio transmitters and receivers require power supplies for the various electrodes. A vacuum tube is normally operated with a-c power on the filament, negative d-c voltage or bias on the control grid, and much greater positive d-c voltage on the plate. Screen grids usually operate at positive d-c voltage, and suppressor grids at ground potential. In some industrial applications, unrectified a.c. is applied to the plate, and the rectifying properties of the vacuum tube are used to eliminate the need for a d-c power supply. But for communication purposes, d-c supplies are necessary because a-c supplies cause noise enough to interfere with intelligible transmission and reception.

A transistor usually operates with a small d-c voltage on the emitter or base and somewhat greater d-c voltage on the collector. Current requirements are small. Transistors are well suited to small portable receivers and to other applications in which space, weight, and power consumption must be small.

Because of the universal use of d-c power in the radio field, this chapter primarily deals with d-c power supplies.

2. A-c and D-c Supply Lines. By far the greater number of transmitters and receivers are operated by power taken from public-utility 60-cycle a-c supply lines. The excellent frequency control of large modern power systems, together with the convenience and comparative freedom from power interruption, makes this source of power convenient for communication apparatus. D-c voltages are mostly supplied by rectifiers. Exceptions are bias voltages derived from plate current flowing through cathode resistors (see 3V4 amplifier filament resistor in Fig. 1), or grid current flowing through *RC* networks. These light circuits are best studied in combination with the amplifiers and oscillators of which they are parts. Rectifiers may be of many kinds, and are taken up in Secs. 9 to 21. In those localities where d-c power is used for all other purposes, it becomes necessary that radio equipment, too, operate from the d-c supply lines. Some receivers, such as that shown in Fig. 1, may be operated from an a-c or a d-c supply line or batteries. Tube plate voltage is limited to that of the d-c source. The rectifier is half-wave because there is no transformer. When operating from a d-c line, the rectifier acts simply as a series resistance, but it protects the (usually electrolytic) filter capacitor if the input polarity is wrong. Dynamotors and m-g sets were formerly used for d-c plate voltages in the range of 250 to 600 volts, but they were heavy, expensive, and inflexible with respect to voltage. With the trend toward smaller over-all size and weight, electronic or vibrator devices have largely supplanted rotating machines for d-c source operation.

3. Isolated Power Plants. In districts isolated from utility-system power lines, engine-generator sets are used for electric-power generation. Engines are commonly of the diesel or gasoline types and range in output from 5 to 500 kw. Choice of engine depends on available fuel. If a reciprocating gas engine is used, the user may choose between gasoline and liquefied gas fuels or between air and liquid-cooled engines, according to local conditions. With rapid variations in connected load, both voltage and frequency fluctuations exceed those of a large power network. Ordinary limits for voltage are ±10 per cent, and for frequency ±5 per cent, from nominal

[1] Westinghouse Electric Corp.

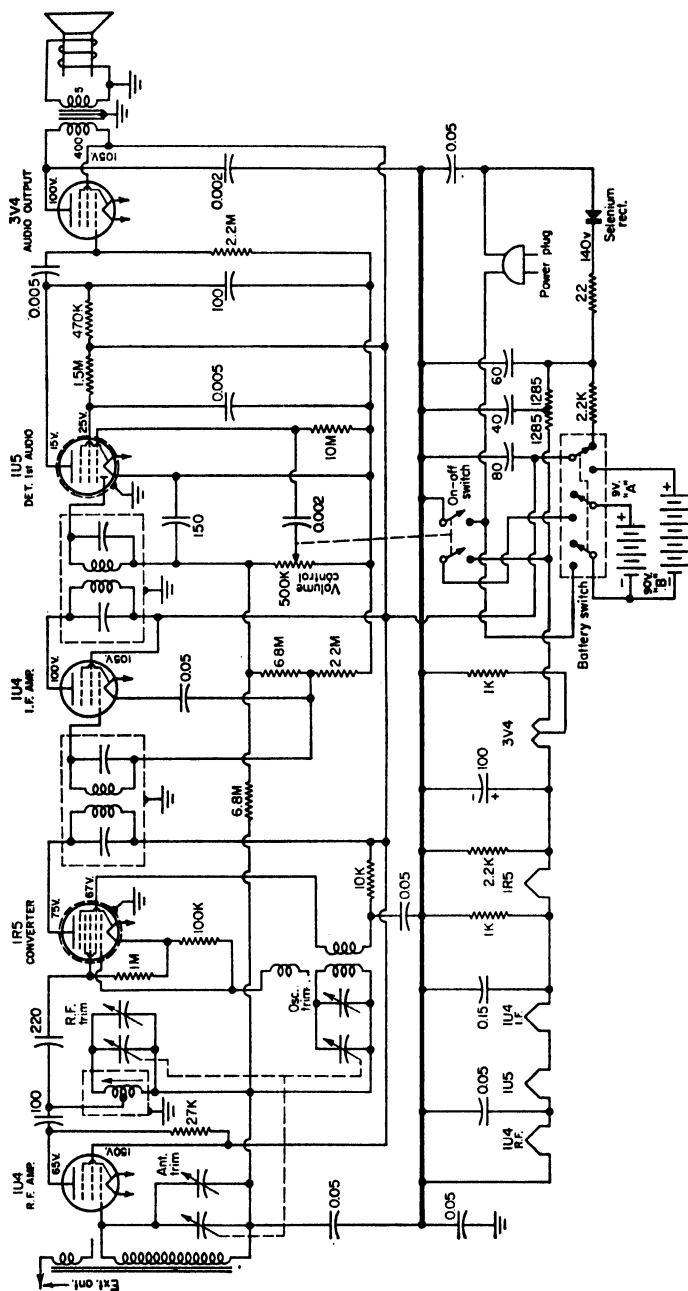



FIG. 1. Schematic of Westinghouse H-342P5U receiver showing arrangements for using battery or a-c power.

ratings, even with fast-acting generator-voltage regulators. Satisfactory operation of communication equipment may then require supplementary voltage regulation, several types of which are described in Secs. 35 to 37.

Voltage and frequency fluctuations are mitigated by the use of engine-speed governors. Under steady-state conditions, the usual mechanical governors provide fair speed control and maintain generator voltage within ±2 per cent. But under

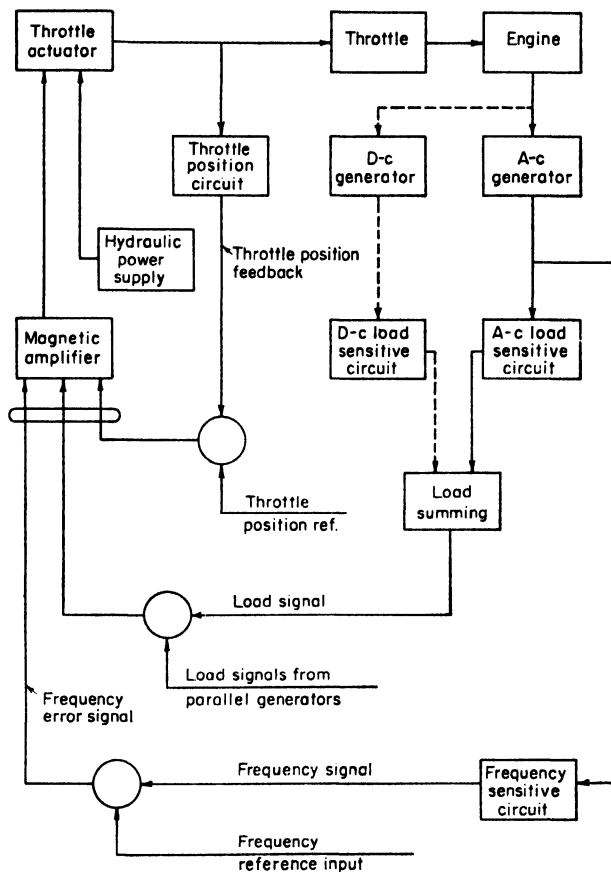



FIG. 2. Electric-generator control circuits.

transient load conditions the generator speed may fluctuate widely. For accurate control under these conditions, hydraulic and electric governors have been developed; the electric governor is the faster of the two.

The Westinghouse electric governor is an accurate speed regulator for engine-generator sets. It is applicable to diesel- and gasoline-engine-driven electric-generator sets from approximately 30 kw capacity to the largest sizes, at 50/60 cps or 400 cps.

It senses electrical load change and thereby anticipates speed change. It is therefore faster acting than a mechanical governor which must wait for actual speed change of the driving engine before starting regulator action.

The governor consists of three major component units: an electrical control unit, a throttle actuator, and a hydraulic power supply. The electrical control unit measures generator frequency, generator load, and engine-throttle position. A block diagram is shown in Fig. 2.[1]

The frequency-sensing circuit is a high-gain low-time-delay loop which provides fine speed control under steady-state conditions and fast response under transient conditions. Its primary function is to counteract small frequency shifts at any one load. The load-sensing circuit anticipates the change in prime-mover speed and repositions the throttle to a new load setting before engine speed has materially changed. At the instant of a load change, the throttle-position-indicating circuit receives the error signal from the load-sensing circuit which causes the throttle to be moved to the final throttle position to hold the new load at base frequency The load-sensing, frequency-sensing, and throttle-position-indicating signals feed into a magnetic amplifier totalizer which in turn passes a signal to the control valve of the hydromechanical actuator.

The hydraulic power supply consists of a hydraulic fluid reservoir that houses a hydraulic pump, filter, relief valve, and necessary line connections; the oil reservoir capacity is 1 gal. The pump is driven by a power take-off on the prime-mover engine and absorbs about 1 hp.

If an increased load is suddenly applied to the engine-generator set, the following action results:

1. The wattmeter output for the new load is instantaneously higher than the throttle-position feedback signal for the previous load.

2. This error signal in the load circuit causes the actuator to open the throttle and hold greater load. The actuator begins to move in approximately 0.015 sec and overshoots the correct throttle position for the new load to force the engine, and then returns to the correct position. As the throttle is opened, the throttle feedback signal increases from the previous value and causes the throttle to stop at the new load position where it again nulls the signal in the load potentiometer.

3. As the throttle approaches the new load position and the load signal is nulled, the frequency-sensing circuit makes a final correction and holds the set on frequency at the new steady-state position. The frequency correction comes into play, since with sudden application of load the speed and frequency drop even though the throttle began to move within 0.015 sec of load application. When the frequency drops, the output of the frequency-sensing circuit produces an error signal which causes the throttle to bring the set back on frequency. Performance obtained is as follows:

Steady-state regulation (no load to full load) ± 0.25 per cent
Transient frequency deviation (no load to full load) 1.5 per cent
Transient recovery time (no load to full load) 1 sec

Electric governors provide for automatic load sharing by two or more engine-generator sets so that the difference between a single generator load and its proportional part of the total combined generator load is within 5 per cent of its rated load. This division of load will be accomplished with ±0.25 per cent frequency regulation. Precise regulation is possible because load sharing is accomplished electrically and, therefore, large speed droop is not necessary. The electric governor works on zero per cent speed droop. Speed droop is adjustable from 0 to −5 per cent by manual adjustment where the electric governor operates with droop in parallel with hydraulic governors.

To realize the foregoing performance, engine response must be rapid, and engine-throttle characteristic (horsepower output vs. throttle position) must be very nearly linear.

4. Stand-by Engine Generators. To provide reliable communications at all times regardless of power-source interruptions, transmitting stations are equipped with stand-by engine-generator sets. A stand-by generator must in itself be reliable, else it fails its purpose.

Engine speed, fuel, cooling-system batteries, battery charger, generator, and starter are chosen to meet specified performance at all extremes of ambient temperature,

[1] Westinghouse Descriptive Bulletin 21000.

humidity, altitude, life, load variations, and power factor. Performance specified in EIA Standard RS-173 on Microwave Standby Generators[1] is

Voltage regulation, no load to full load, ±5 per cent over-all
Frequency regulation 20 per cent load to full load, ±1 cycle
Wave shape, within 10 per cent of a sine wave at all points

Full engine speed should be attained within 15 sec after starting, should not vary more than 5 per cent during a load change, and should return to full speed within 8 sec after a load change.

In unattended stations an automatic-transfer panel is provided to switch over to stand-by power whenever the line voltage drops below 98 volts for more than 5 sec. If the stand-by power unit fails, or if line voltage is restored to 104 volts for a period of at least 10 min, the load switches back to the line automatically. Figures 3 and 4 show recommended installation of a stand-by gas engine-generator set.

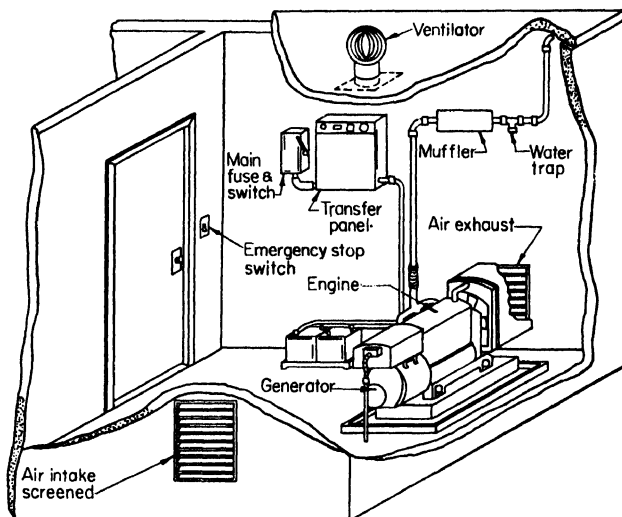



FIG. 3. Stand-by generator installation.

Maintenance of the stand-by power plant is similar to that familiar to automobile owners, but to provide good reliability it should be scheduled and systematic. The engine should be started periodically. By means of a momentary failure-simulating test switch, the load should be transferred to the engine-driven generator, and meter readings recorded. The load should be left on the generator for a period of 20 min to dry out the lubrication system. At the end of this period, the test switch should be released. Automatic transfer of load back to the power line then takes place.

Battery-charging rate must be adjustable and automatically regulated to replace normal leakage and occasional starting drain. Overcharging may easily occur, with resulting fast disintegration of the positive grids.

5. Battery Operation. Batteries for radio applications are of either the dry or storage type. Dry batteries are used in small portable receivers and instruments. In the design of these devices, precautions should be taken not to shorten the battery life by excessive current drain. Dry-battery shelf life is best at low ambient temperatures. The battery should be replaced before terminal voltage falls to the point at which faulty operation begins. Storage batteries are used in transportation

[1] Electronic Industries Association Standard RS-173, "Emergency Standby Power Generator and Accessories for Microwave Systems." See also Tharp, N. B., Recommendations for Improving Reliability of Standby Engine Generators for Microwave Communications Systems, *Trans. AIEE*, III, 261, April, 1955.

apparatus: automobiles, buses, trucks, airplanes, and trains. Batteries are also used where it is vital that communication and control facilities function without interruption when line power fails. Examples are central power stations and substations. Here the battery voltage is 110 to 125 volts. Care of the batteries is an important part of plant maintenance and should be done in accordance with the battery manufacturer's recommendations. Such is the emphasis on continuity of service that batteries are used at all times, even though station a-c power is available except during very infrequent shutdowns. In power-line carrier telephony, automatic simplex operation requires that the transfer from "send" to "receive" be accomplished in 0.001 sec. This requirement alone eliminates the possibility of switching to any type of stand-by power. Vacuum tubes must operate at the battery voltage, and this restricts the choice of tubes as well as the power output. Such tubes as the 25L6 are extensively used in such service.

Fig. 4. Fuel system for stand-by generator.

6. Vibrators. Vibrators constitute an important means of deriving plate and screen potentials from a storage battery, especially in transportation apparatus. Typical battery voltages are:

6 volts in automobiles
12 volts in buses, trucks, and aircraft
24 volts in aircraft and boats
32 volts in boats, trains, and trolley cars
64 to 110 volts on trains

Vibrators convert these voltages into plate potentials of 250 to 500 volts.

The reliability and performance of vibrators have improved vastly during recent years, especially in the so-called "heavy-duty" size. The vibrator is essentially a vibrating metal reed with a frequency of vibration determined by the mass and stiffness of the reed. Low-power (30-watt) vibrators operate at 115 to 250 cps; heavy-duty (150-watt) vibrators at 60 cps. With good design, the contacts close with very low relative velocity and open with comparatively high velocity. This action eliminates contact bounce, reduces sparking at the "break," and reduces wiping motion and contact erosion. Two principal kinds are used, the self-rectifying and a-c output vibrators. The self-rectifying vibrator with shunt coil is shown in Fig. 5.[1] Current in the coil attracts the reed armature and

Fig. 5. Self-rectifying vibrator with shunt coil.

[1] P. R. Mallory and Co., "Fundamental Principles of Vibrator Power Supply Design," 1947.

starts the vibrations. Current flows first through one half of the primary and then through the other as the vibrator reed switches from left to right contacts. Simultaneously the secondary circuit is open and closed and thus provides d-c output of a voltage determined by the turns ratio of the transformer. A timing capacitor is connected across the transformer primary or secondary to reduce arcing and maintain good wave form.

FIG. 6. Vibrator with separate driver and a-c output.

During the time that the vibrator is accelerating to its normal speed, heavy transient currents flow through the contacts and, unless provisions are made to protect against them, may easily burn up the contacts in a short time. One method of protecting contacts during the starting interval is the so-called "separate driver circuit" of Fig. 6. Here all the power contacts are separated when the reed is at rest but the driver contacts are closed. Power contacts are made of tungsten and driver contacts of precious metal. Because the moving driver contact causes an unbalanced mechanical system careful design is necessary to produce uniform wave shape on both halves of the output wave. Transformer input and output wave forms are shown for correctly adjusted vibrators of the self-rectifying type in Fig. 7, and for the a-c type in Fig. 8. In Fig. 7, separation of contacts starts a transient oscillation as shown by the dotted line. At the end of the contact transfer interval t_2, the other contact closes and damps out the transient. The same thing occurs when the other contacts open and close in the reverse direction during interval t_4. In an a-c vibrator, the contacts reverse during the time when the transient wave crosses the zero axis, thus giving smooth a-c output. A separate rectifier is then used for d-c output.

FIG. 7. Input and output wave forms of self-rectifying vibrator.

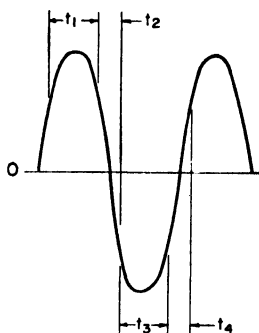



FIG. 8. Input and output wave form of a-c vibrator.

It will be noted that the self-rectifying type of vibrator produces d.c. with deep but narrow notches of ripple. This ripple can be smoothed out with a small filter. The effects of contact chatter and bounce, and too much or too little timing capacitance, are shown in Fig. 9.

FIG. 9. Vibrator wave forms with improper conditions.

There is always a small amount of sparking at vibrator contacts. This is a source of radio interference, and the case, shielding, and grounding connections must be arranged carefully, and an r-f filter added to minimize this interference. The audible noise from a vibrator is usually made tolerable by means of cushioning material glued to the inside of the case. The arcing that may occur if one vibrator contact is not making or breaking properly is serious and results in wearing out contacts rapidly. It may be detected by observing the wave shape and may be corrected by contact adjustment.

Arcing of this kind can be detected as a flame in contrast to the tiny sparks of normal operation.

In Fig. 10 the elementary components of a sine-wave a-c 150-watt power supply are shown.[1] Although this circuit shows a vibrator *V* with one power-handling circuit, actually there are four parallel contacts on each side. The vibrator has a separate driving circuit and arc-suppression network made up of C_d and *R*.

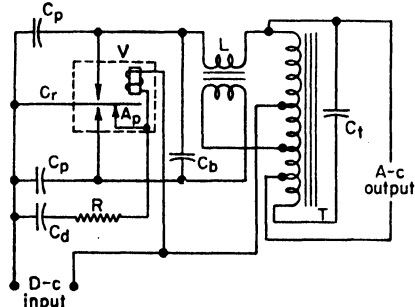



FIG. 10. Basic circuit of heavy-duty vibrator.

When rated d-c voltage is applied, with the vibrator at rest, coil current is drawn from the power source through the center reed C_r and the actuating point A_p. This causes the coil to attract the vibrator armature and start the reed vibrating. The nominal frequency is 60 cps. The vibrator is connected through equal and oppositely phased reactor windings *L* to a center-tapped transformer *T*.

Buffer capacitor C_b connected across opposite sides of the vibrator, in conjunction with the reactor *L*, transformer *T*, and tank capacitor C_t, forms a roughly tuned circuit which reverses the direction of current in the transformer primary during the interval when the vibrator center reed swings from one side to the opposite side. This reversal of current prevents heavy arcing at the contacts which would be present if they were required to break the maximum value of direct current which is built up in the primary winding.

Balanced reactor *L* is of a linear type and works in conjunction with the saturated transformer and tank capacitor to produce an essentially sinusoidal output wave. Autotransformer *T* in this simplified circuit has, in addition to its center-tapped primary winding, an extension of sufficient turns to produce the output a-c voltage. Tank capacitor C_t is connected across this extension. On this supply the tank capacitor operates at approximately 330 volts.

Power-point capacitors C_p are connected between the vibrator center reed and the stationary contacts. They are used primarily to maintain stability of operation during the starting interval. Their function is to prevent a flare arc at the contacts when the vibrator is coming up to speed from rest, and while the vibrator frequency is below the resonance frequency of C_b, *T*, and C_t.

When the vibrator approaches the end of useful life, the tungsten material on the contacts of the vibrator is gradually used up. This causes sticking of the contacts, or a flare arc, which causes excessive primary current to be drawn and a circuit breaker to trip. It is then necessary to replace the vibrator. It is usually possible to determine the degree of deterioration of vibrator contacts in advance of actual failure by noting the voltage at any output and comparing it with the proper value for any given condition of input voltage. The gradual loss of tungsten at contacts will cause the time efficiency of the vibrator to be reduced. For a new vibrator, the total on-time interval for each contact is approximately 40 per cent of one cycle. Considering both sides of a set of power contacts, this is an 80 per cent time efficiency. With rated d-c input, this time efficiency will give, say, 300 volts d.c. at full load. When this voltage deteriorates by as much as 15 per cent, the contacts have worn to the point where time efficiency is 69 per cent instead of 80 per cent. When this amount of wear shows up, it is quite likely that at least some of the contacts are approaching a worn-out condition, and it would be well to replace the vibrator. Output voltage in the power supply, under load, can be considered as only an indication of the average condition of the contacts. All contacts are not set exactly the same; hence it is possible for one or more contact pairs to fail even though the average is still acceptable. This condition can be found by inspection.

[1] Cornell-Dubilier Electric Corp. Service Manual 3600, "Railroad Vibrator Converter," 1955.

7. Aircraft Power Supplies. Aircraft electric power is supplied either from storage batteries or from engine-driven generators. In the larger planes, a-c generators are rated up to 60 kva, three-phase, 208 volts, 400 cycles with 120 volts to neutral. Generators driven directly by airplane engines may vary widely in speed and therefore in frequency. Hydraulic transmissions are interposed between alternators and engines to provide constant-speed drives. In airplanes equipped with separate air turbines to drive the generators, the turbines are controlled hydraulically to give constant speed. In both the constant-speed systems, standard variations are ±5 per cent for both voltage and frequency; in other systems these variations may be much wider.

Some electronic equipment requires more precise frequency and voltage regulation. With magnetic amplifier control of the generator voltage regulator and of the transmission-control devices, frequency can be stabilized to 400 cps ±1 cycle, and voltage to 200 ± 5 volts. This has been accomplished with two alternators operating in parallel, and with provision for rapidly switching the load to a ground power unit at times of engine shutdown.[1]

Aircraft-generator wave form varies appreciably between units and from no-load to full-load conditions. This affects d-c power-supply voltage. The 400-cycle supply frequency is subject to random or cyclic low-frequency modulation (2 to 70 cps) which may interfere with amplifier operation and which would require prohibitively large components to be filtered out of the power supply. Performance specifications generally limit this modulation to 2 per cent.

CHARACTERISTICS AND MEASUREMENT OF D.C.

8. Indicating Instruments. Since d.c. is employed largely for radio transmitters and receivers, a brief analysis will be made of its various kinds and their measurement. One reason for this analysis is that instruments of the repulsion-iron or dynamometer type will not read the same as an "average" type on certain kinds of d.c. This difference sometimes causes confusion.

If a d-c ammeter is specified, it usually refers to a D'Arsonval instrument (permanent magnet, moving coil), one that reads "average" values.

Figure 11 shows five typical kinds of d.c., one or more of which are present in any d-c power supply. Steady d.c. is the output from a primary or secondary (storage) battery, or from a suitable filter connected to a pulsating or ripple d-c source. Ripple d.c. is the usual output from a d-c generator, the ripple being a result of commutation. Three types of pulsating d.c. are (1) half-wave rectified single-phase, (2) full-wave rectified single-phase, and (3) three-phase full-wave rectified.

The ambiguity of the term "direct current" is readily apparent when considering Fig. 11, since all these wave forms fall into this classification; *i.e.*, each remains positive. None goes to a negative value. The data on the figure show the minimum voltage as a decimal part of the maximum voltage E_m, *e.g.*, 65 per cent of its maximum, for ripple d.c.

The second column shows the average value of the potential difference as a factor times the maximum E_m. The factor for pulsating d.c. varies from 0.32 for the half-wave rectified single-phase to 0.96 in the case of the full-wave three-phase.

The rms or effective value of a current is such that the heating effect (I^2R) is the same for d.c. or a.c., by definition. For pulsating d.c. the wattage found by multiplying the average voltage by the average current is not the same as the rms voltage times the rms current. The results of these average readings are sometimes called "d-c watts."

The readings of different types of instruments can be predicted from the value of the form factor. On the ripple d.c. from a d-c generator, when the form factor is 1.01 as shown, a D'Arsonval instrument would read 1 per cent lower than would a dynamometer type of instrument. The D'Arsonval instrument reads average, and the dynamometer reads rms. When there is a difference in readings, the rms instru-

[1] Oman, L. R., "An Automatic A-C Generating System for Aircraft," AIEE Paper DP 56-462.

ment always reads higher. For the pulsating d-c output of the single-phase half-wave rectifier, the form factor is 1.57, and an rms instrument (repulsion-iron, dynamometer, or thermocouple) would read 57 per cent higher than the average-reading instrument (D'Arsonval).

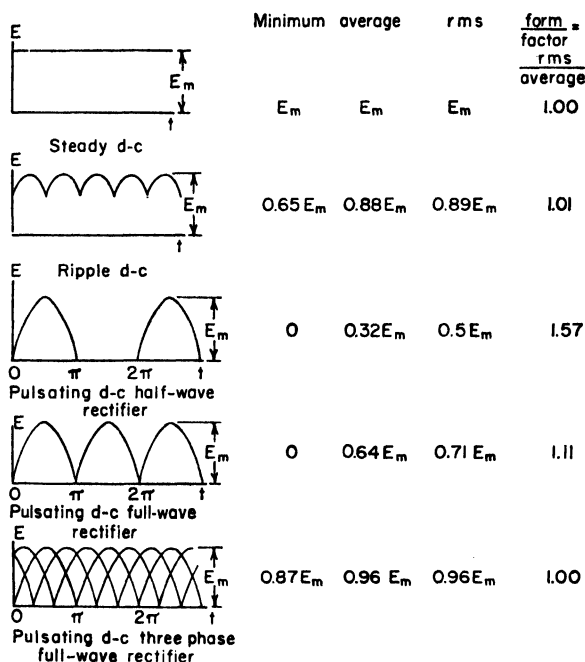



Fig. 11. Types and characteristics of d-c power.

Power transformers for radio receivers and amplifiers may be described by specifying d-c volts and d-c amperes output. However, to design a transformer which has rms a.c. flowing in its windings to supply a filter which furnishes average d.c. to the set, the form factor is employed. A full-wave rectifier has the characteristic form factor 1.11, which means that this is the ratio of the rms to average. To supply 100 ma average, the transformer furnishes 111 ma rms (55 ma from each secondary). See Table 1 on page 17-11.

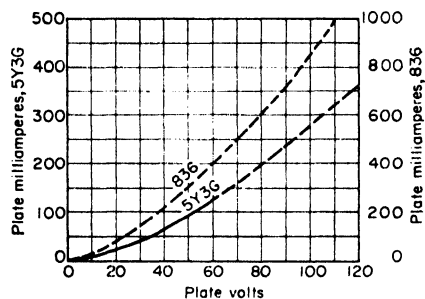



Fig. 12. Characteristics 836 and 5Y3G rectifiers.

RECTIFIER POWER SUPPLIES

9. Rectifier Tubes. Rectifiers are used to convert alternating into direct current. Rectifier tubes or diodes have two electrodes (cathode and anode) and are either high-vacuum or gas-filled. Both kinds are available in a wide range of voltages and currents, from miniature tubes suitable for fractional-watt output to large ones for hundreds of kilowatts. Vacuum rectifiers are characterized by comparatively large internal voltage during conduction known as the "plate characteristic," as shown in Fig. 12. These are for types (1)

Table 1. Commonly Used Rectifier Circuits

TYPE	SINGLE PHASE HALF WAVE	SINGLE PHASE FULL WAVE	SINGLE PHASE BRIDGE CIRCUIT	3 PHASE HALF WAVE	3 PHASE HALF WAVE ZIG ZAG	DOUBLE 3 PHASE WITH BALANCE COIL *	3 PHASE FULL WAVE (SEC. MAY BE Δ)	3 PHASE FULL WAVE ZIG ZAG #	6 PHASE HALF WAVE
CIRCUITS	+ − SEC. PRI.	+ −	+ −	+ −	+ −	+ − BALANCE COIL	+ −	$\frac{+HV}{2}$ − +HV	− +
RECTIFIER PHASES AND NUMBER OF TUBES	1	2	2-4	3	3	6	6	6	6
PHASES OF A-C SUPPLY	1	1	1	3	3	3	3	3	3
SECONDARY VOLT PER LEG	2.22	1.11 (HALFSECT)	1.11 (WHOLE)	0.855	0.985 (HALF LEG .493)	0.855	0.428	0.493 (HALF LEG .247)	0.74
PRIMARY VOLTAGE	2.22	1.11	1.11	0.855	0.855	0.855	0.428	0.428	0.74
SECONDARY CURRENT PER LEG	0 1.57	0 0.707	0 1.00	0 0.577	0 0.577	0 0.289	0 0.816	0 0.816	0 0.408
PRIMARY CURRENT PER LEG	0 1.21	0 1.000	0 1.00	0 0.471	0 0.408	0 0.408	0 0.816	0 0.816	0 0.577
SECONDARY KVA	3.48	1.57	1.11	1.48	1.71	1.48	1.05	1.21	1.81
PRIMARY KVA	2.68	1.11	1.11	1.21	1.05	1.05	1.05	1.05	1.28
AVERAGE OF PRIMARY AND SECONDARY KVA ø	3.08	1.34	1.11	1.35	1.38	1.26	1.05	1.13	1.55
INVERSE PEAK VOLTAGE	3.14	3.14	1.57	2.09	2.09	2.09	1.05	1.05	2.09
RMS CURRENT PER TUBE	1.57	0.707	0.707	0.577	0.577	0.289	0.577	0.577	0.408
PEAK CURRENT PER TUBE	3.14	1.00	1.00	1.00	1.00	0.500	1.00	1.00	1.00
AVERAGE CURRENT PER TUBE	1.00	0.50	0.50	0.33	0.33	0.167	0.33	0.33	0.167
RIPPLE FREQUENCY	f	2f	2f	3f	3f	6f	6f	6f	6f
RMS RIPPLE VOLTAGE	1.11	0.472	0.472	0.177	0.177	0.04	0.04	0.04	0.042
RIPPLE PEAKS	+2.14 −1.00	+0.57 −1.00	+0.57 −1.00	+0.209 −0.291	+0.209 −0.291	+0.057 −0.077	+0.057 −0.077	+0.057 −0.077	+0.057 −0.077
LINE POWER FACTOR	0.373	0.90	0.90	0.826	0.955	0.955	0.955	0.955	0.955

NOTES: The values of voltage and current are effective or RMS, unless otherwise stated; they are given in terms of the average D-C values, and the KVA in terms of D-C kw. output. Perfect transformers, rectifiers and D-C choke are assumed, and $N_1/N_2 = 1$, except in zig-zag circuits. Secondary and primary refer to anode transformer.

* Frequency of balance coil voltage 3f. Balance coil voltage 0.356. Peak balance coil voltage 0.605. Balance coil KVA 0.173.

\# Use column 5 values for half voltage tap. ø Magnetizing current assumed negligible in all cases.

5Y3G and (2) 836, with peak inverse voltages of 1,400 and 5,000, respectively. These curves are shown solid up to the average current ratings.

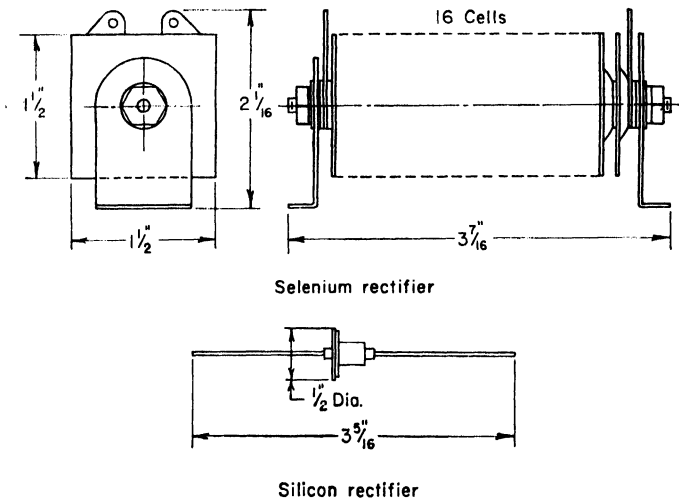
Plate-characteristic voltage divided by plate current gives effective tube resistance at any point. Tube resistance decreases as current increases, up to the emission limit, where all the electrons available from the cathode are used. Filament voltage governs the emission limit and must be closely controlled. If the filament voltage is too high, the tube life is shortened; if too low, the tube will not deliver rated current at the proper voltage.

	Selenium	Silicon
Average current, ma	150	200
Inverse peak voltage	560	500
Forward drop, volts	20	2
Reverse current, ma	10	0.002
Max ambient temp, °C	35	100
Weight, lb	0.5	0.002
Max operating frequency, cps	2,000	20,000

FIG. 13. Relative dimensions of selenium and silicon rectifiers.

Indirectly heated cathodes have a filament inside a mass of ceramic material coated with the emitting material. These cathodes reduce hum and make quieter operation possible. The thermal time constant of the cathode is approximately 30 sec.

Gas-filled rectifier tubes have an internal voltage drop which is virtually constant and independent of current. Usually this voltage drop is much lower than that of high-vacuum tubes. Consequently, gas-filled tubes are used in high-power rectifiers, where high efficiency and good regulation are important. Gas tubes are limited as to operating frequency. Usually this frequency is below 150 cps but for some (type 866) may be as high as 1,000 cps. These tubes are limited to moderate ambient temperatures (20 to 70°C). Gas-filled tubes also have two current ratings; peak and average. For example, the 872 tube is rated at 5 amp peak and 1.25 amp average plate current. In this tube the internal voltage drop is 10 volts and the peak inverse voltage is 10,000. This is the voltage which occurs in the nonconducting period of the

rectifier. Large gas-filled rectifiers sometimes have tube-filament voltages in quadrature to reduce hum. In high-voltage operation these tubes are subject to occasional arc-back, or conduction in the reverse direction. Rectifier-tube life is increased if cathodes are allowed to reach full operating temperature before plate voltage is applied.

10. Solid-state Rectifiers. Metal-junction rectifiers exhibit greater conduction in the so-called reverse direction, but internal voltage drop is less than in high-vacuum rectifiers. Metal rectifiers are generally smaller but are more limited in temperature and frequency ranges.

Selenium rectifiers are used in both receivers and transmitters. These rectifiers are limited in top operating temperature and for high power are generally air-cooled. Current ratings are increased by means of radiating fins. Unlike vacuum diodes, selenium rectifiers have appreciable reverse current flow, which is a direct loss in efficiency and power output. Usual practice limits the voltage per cell to 35. Internal voltage during conduction is called "forward drop." In humid climates, selenium rectifiers must be protected against penetration of atmospheric moisture to maintain good rectifying properties. In power-supply sizes, operation is limited to frequencies below 2,000 cps.

An improvement over selenium for low and medium power is the silicon diode.[1] This rectifier has small dimensions, will operate at high temperatures and frequencies, and has low forward drop and reverse current. Single diodes have peak inverse voltages up to 1,000. Silicon diodes have the characteristic property that at maximum inverse voltage (Zener voltage) the voltage curve bends abruptly downward with extremely steep slope. Within the limits imposed by reverse current dissipation, this property makes silicon rectifiers useful as voltage regulators. Figure 13 indicates the relative dimensions and properties of selenium and silicon rectifiers. Silicon-rectifier leads should be maintained at the length indicated in Fig. 13 but may be bent to save space. Higher power is obtained by mounting the rectifier on an aluminum plate. Figure 14 shows a silicon cell arranged for plate mounting, and Fig. 15 gives the current ratings for this rectifier with two plate sizes.

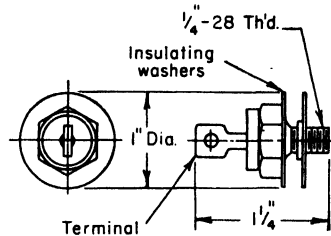



FIG. 14. Stud-mounted silicon power rectifier.

Germanium rectifiers are intermediate between selenium and silicon in some respects. Dimensions are approximately the same as the silicon diode of Fig. 13. Forward drop and peak inverse voltage are somewhat lower, and reverse current higher, than in silicon diodes of comparable current rating. The severest limitation is that of operating temperature. Germanium rectifiers are rated for a maximum operating temperature of 65°C. At this temperature, maximum peak inverse voltage and forward current drop to as low as 30 per cent of the 25°C ratings.

Both silicon and germanium diodes are hermetically sealed against ambient moisture and are often mounted on metal plates to increase current ratings. Figure 16 is a curve showing relative performance of selenium, silicon, and germanium rectifiers at 55°C.

11. Thyratrons and Ignitrons. If a grid is placed between anode and cathode in a hot-cathode gas-filled rectifier tube, the tube is called a thyratron. Rectified output may be controlled by means of a-c or d-c voltage applied to thyratron grids. A voltage greater than "critical" must be applied to permit conduction; but after conduction starts, reduction in grid voltage does not affect conduction. This action produces a

[1] James, H. B., "Design and Development of a New Silicon Rectifier Aircraft D-C Power Supply," AIEE Paper DP 56-447. Perlman, Sol, The Power Supply in Military Equipment, *IRE Convention Record*, Part 6, 1956, discusses the advantages in decreased weight, complexity, and heat problems gained from the use of metal-rectifier power-supply systems.

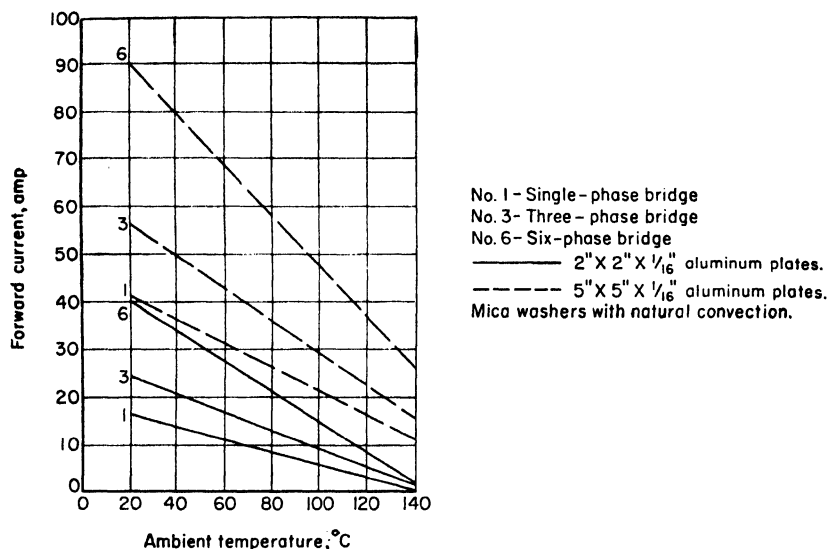



FIG. 15. Silicon power rectifier current ratings—peak inverse voltage 200 volts.

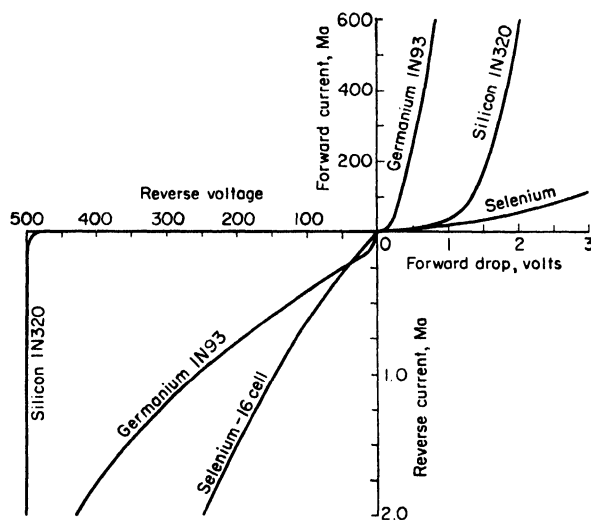



FIG. 16. Comparison of semiconductor rectifiers.

high peak-to-average load-current ratio. Thyratrons are manufactured in most of the gas-filled rectifier-tube sizes, with ratings up to 15,000 volts peak inverse and 6 amp average plate current. They are liable to arc-back under the same conditions as for two-anode gas-filled rectifiers.

For higher output current at power frequencies, ignitrons are used. An ignitron has a cold cathode comprising a pool of mercury. Control is obtained by an ignitor electrode which projects into the mercury pool and which, if supplied with a relatively small electric current, starts the conduction each cycle as desired. The ignitron is

formed chiefly of metal except for the glass insulation which supports the graphite anode in the upper part and the ignitor in the lower end of the tube. The conduction is chiefly by electrons moving from the arc spot on the pool to the anode. The negative electron space charge is neutralized by the formation of positive ions in the mercury vapor, thus allowing the passage of large currents with small potential drop. Use of a metal envelope and water cooling allows high current ratings for relatively small-sized tubes.

Most ignitrons have a water jacket made of stainless steel and provided with spiral baffles built in as a part of the tube. The smallest size is designed to be installed in a clamp which also cools the tube by air or water cooling.

Ignitrons designed for rectifier service are required to operate once each cycle continuously, hence are constructed with baffles within the envelope to shorten the deionization time and thereby minimize the possibility of arc-back. They are provided with two ignitors, one of which is a spare and either of which is suitable for continuous operation on the various circuits used in rectification service, as well as for services where the ignitor is operated by connection to the main anode through a thyratron.

An auxiliary anode, to which the ignitor current is transferred when the tube is fired, is provided to make it possible to prolong the excitation period in circuits where separate ignitor excitation is used. This extension of ignition time is necessary where interphase circuits are used with separate excitation. It is not advisable to operate the separate ignitor-excitation circuits for more than 1 min before starting conduction on the main anode as this may cause mercury to condense on the main anode, increasing the danger of arc-back.

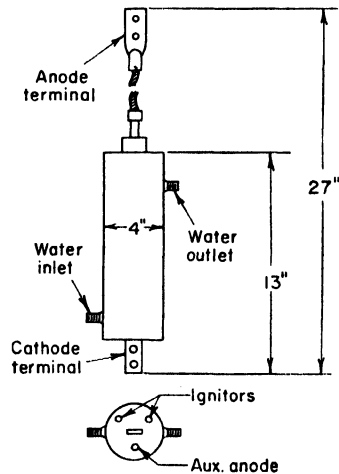
Fig. 17. Dimensions of 5554/679 ignitron.

These tubes must be mounted vertically with the cathode pool down. Operation will be most stable if they are protected from shock and vibration. When the anode of the ignitron is +50 volts or more with respect to the cathode, and a relatively small current is passed through the ignitor into the mercury pool (the ignitor positive with respect to the pool), an arc spot will form on the pool close to the ignitor-mercury junction and conduction to the anode is immediately established.

When ionization occurs the tube will remain conducting until the anode current drops to zero. It is thus necessary to ignite or fire the ignitron in each cycle that operation is desired. Voltage can be applied to the anode simultaneously with its application to the control electrode or ignitor.

After the ignitor has produced the initial arc, no further current need flow through it, and if current is allowed to flow, it will produce overheating of the ignitor. For this reason, circuits should be used which will furnish current to the ignitor in pulses. Firing or ignition of the ignitrons can be accomplished at any point in the a-c cycle by a timing circuit using a thyratron connected between the anode and ignitor of each ignitron. By this means it is possible to control both the number of cycles during which the ignitron operates and the effective current per cycle.

The frequency at which an ignitron operates is limited by the speed of deionization during the half cycle when the tube is not conducting and depends very greatly on the water temperature and anode current. Hence ratings change whenever it is desired to operate the tube at frequencies outside the range of 25 to 60 cycles. The 5554/679 ignitron has dimensions shown in Fig. 17 and 60-cycle ratings of 75 amp average, 600 amp peak, and 4,500 amp surge currents. The peak inverse voltage is 2,100.

FIG. 18. Chart of selection of a particular rectifier for a particular application. (*Amperex Electronics Corp.*)

The ignitor requires 40-amp 100-μsec pulses at 150 positive peak volts; 1.5 gal per min of cooling water is necessary.

12. Rectifier Voltage, Current, and Circuits. Rectifiers are used for d-c power-supply voltages ranging from less than 200 volts in receivers to over 200,000 volts in klystron transmitters. Transmitter plate voltages are often measured in thousands of volts. Danger of electrocution makes advisable the posting of notices such as the following near transmitters.

> H.V. WARNING: Operation of this equipment involves voltages which are dangerous to human life. Operators should observe all safety precautions at all times. Use grounding stick to discharge all high-voltage terminals before servicing. Use one hand only when making internal adjustments.

Current and voltage ratings of power rectifiers are given in the chart of Fig. 18. Rectifier-current values range from a few microamperes in oscilloscope anode d-c supplies to hundreds of amperes in very large transmitters. Generally the rectifying-device average current differs from the total output current of the rectifier circuit. Table 1 gives commonly used rectifier circuits, together with current and voltage relations in the associated transformers. This table is based on the use of a reactor-input filter to reduce ripple. The inductance of the choke is assumed to be great enough to keep the output d.c. constant. With any finite inductance there is always some superposed ripple current which is neglected in the table; this is considered further in Sec. 13.

Single-phase bridge and three-phase full-wave rectifiers require notably low a-c voltage for a given d-c output and low inverse peak voltage on the tubes. The single-phase half-wave rectifier ordinarily has discontinuous output current, and its output voltage is therefore highly dependent upon the inductance of the input filter choke. For this reason, currents and voltages are given for this rectifier in Table 1 without a filter. Unbalanced d.c. in half-wave rectifiers requires the use of larger transformers than in full-wave rectifiers. This is partly overcome in three-phase transformers by the use of zigzag connections. The three-phase full-wave rectifier can be delta-connected on both primary and secondary if desired; the secondary current is multiplied by 0.577 and the secondary voltage by 1.732. Anode windings have more turns of smaller wire in the delta connection.

13. Ripple.[1] Filters used with rectifiers allow the rectified d.c. to pass to the load without appreciable loss, but ripple in the rectified output is attenuated to the point where it is not objectionable. Different types of rectifiers have differing output voltage waves, as indicated in Table 1. Certain assumptions will be made to simplify this discussion:

1. The alternating voltage to be rectified is a sine wave.
2. The rectifying device passes current in one direction but prevents any current flow in the other direction.
3. Transformer and rectifier voltage drops are negligibly small.
4. Filter capacitor and reactor losses are negligible.

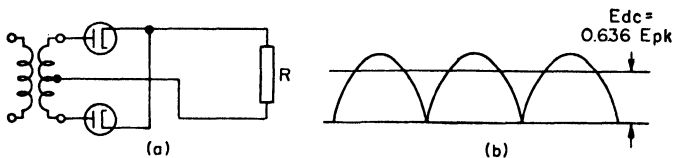



FIG. 19. Circuit and wave form of full-wave rectifier.

A single-phase full-wave rectifier is shown in Fig. 19. The alternating components of the output voltage have a fundamental frequency double that of the supply.

[1] Several of the following sections dealing with filtering of power supplies and with small transformer design are largely from Reuben Lee, "Electronic Transformers and Circuits," 2d ed., Wiley, 1955.

Current flow is continuous and E_{dc} has 63.6 per cent of peak amplitude. A full-wave rectifier uses only one-half of the transformer winding at a time; that is, E_{ac} is only half the transformer secondary voltage. A single-phase bridge rectifier (Fig. 20) utilizes the whole of this voltage in producing E_{dc}. The output voltage relations are the same as those of Fig. 19. Although this circuit requires more rectifying tubes, it does not need a transformer midtap.

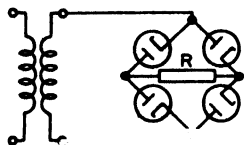



FIG. 20. Bridge rectifier.

The effect of rectifying more than one phase is to superpose more voltages of the same peak value but in different time relation to each other. Figure 21 gives a comparative picture of the rectified output voltage for three-phase half-wave and full-wave rectifiers. Increasing the number of phases increases the value of E_{dc} and increases the frequency and decreases the amplitude of the alternating components. Ripple frequency is p times that of the unrectified alternating voltage, p being 1, 2, 3, and 6 for the respective waves. Roughly speaking, p may be taken to represent the number of phases, provided that allowance is made for the type of circuit, as in Fig. 22. Rectifiers with $p = 3$ or 6 are derived from three-phase supply lines, and,

FIG. 21. (a) Three-phase half-wave rectifier; (b) three-phase, full-wave or double-Y six-phase rectifiers.

by special connections, rectifiers with $p = 9$, 12, or more are obtained. The frequency of any ripple harmonic is mp, where m is the order of the harmonic.

Ripple voltage for any of these rectifiers can be found from

$$A_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos n\omega t \, dt \tag{1}$$

where A_n = amplitude of nth ripple harmonic
T = ripple fundamental period
t = time (with peak of rectified wave at $t = 0$)
$\omega = 2\pi/Tp = 2\pi \times$ supply-line frequency
$f(t)$ = ripple as a function of time
$= E_{pk} \cos \omega t, \; T/2 > \omega t > -T/2$

The voltage peak is chosen at $t = 0$ to obtain a symmetrical function $f(t)$ and eliminate a second set of harmonic terms like those in Eq. (1), but with $\sin n\omega t$ under the integral.

Ripple amplitude is given in Fig. 22 for the ripple fundamental and second and third harmonics with reactor-input filters. The ratio P_A of ripple amplitude to direct output voltage is plotted against the number of phases p. Note that P_A diminishes by a considerable amount for the second and third harmonics. In general, if a filter effectively reduces the percentage of fundamental ripple across the load, the harmonics may be considered negligibly small.

14. Inductor-input Filters. In the inductor-input filter shown in Fig. 23a, the rectifier is a source of nonsinusoidal alternating voltage connected across the filter. For any harmonic, say the nth, the voltage across the whole circuit is the harmonic amplitude A_n, and the voltage across the load is $P_R E_{dc}$, P_R being allowable ripple across the load, expressed as a fraction of the average voltage. Since the load resistance R is high compared with X_C, the two voltages are nearly in phase, and they bear the

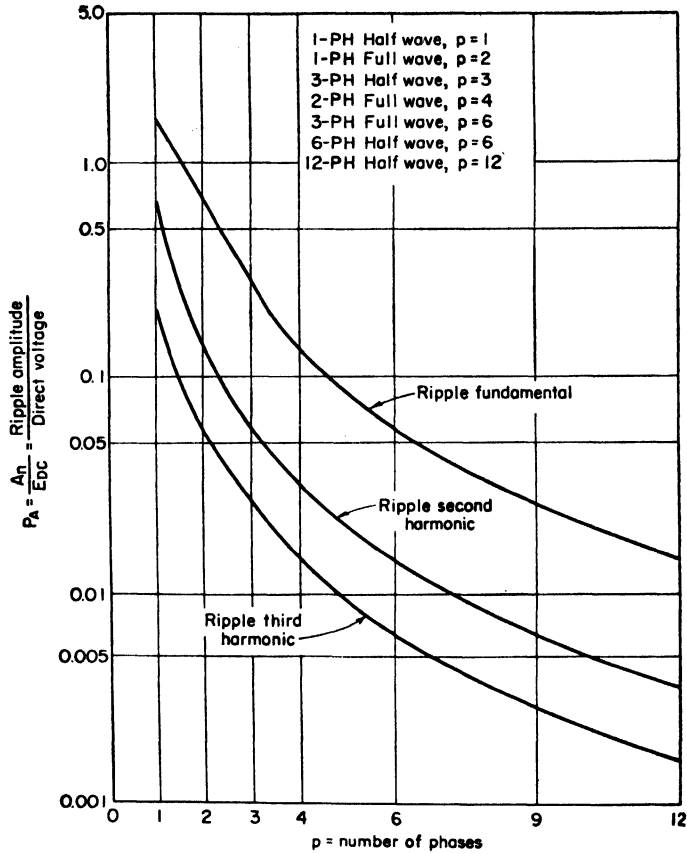



FIG. 22. Ripple content of output of various rectifiers.

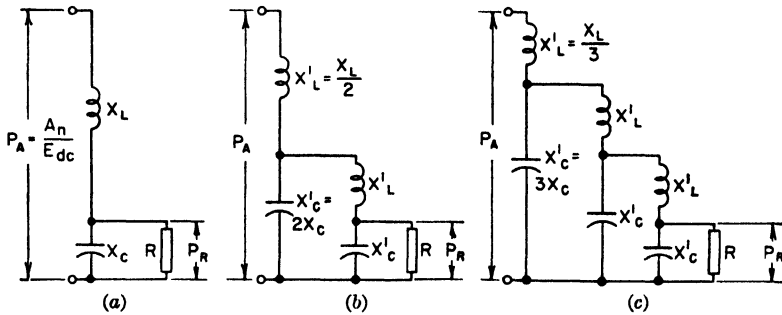



FIG. 23. Inductor-input filter circuits.

same ratio to each other as their respective reactances, or

$$\frac{P_A}{P_R} = \frac{X_L - X_C}{X_C} = \frac{X_L}{X_C} - 1 \tag{2}$$

From the type of rectifier to be used and the permissible amount of ripple in the load voltage, it is possible to determine the ratio of inductive to capacitive reactance.

When P_R must be kept very small, the single-stage filter of Fig. 23*a* may require the inductor and the capacitor to be abnormally large. It is preferable under this condition to split both the inductor and the capacitor into two separate equal units and connect them like the two-stage filter of Fig. 23*b*. A much smaller total amount of L and C will then be necessary. For this filter

$$\frac{P_A}{P_R} = \left(\frac{X_L' - X_C'}{X_C'}\right)^2 \tag{3}$$

X_L' and X_C' being the reactances of each inductor and capacitor in the circuit. The three-stage filter of Fig. 23*c* may be more practicable for still smaller values of P_R. Here

$$\frac{P_A}{P_R} = \left(\frac{X_L' - X_C'}{X_C'}\right)^3 \tag{4}$$

and, in general, for an n-stage filter,

$$\frac{P_A}{P_R} = \left(\frac{X_L' - X_C'}{X_C'}\right)^n \tag{5}$$

It is advantageous to use more than one stage only if the ratio P_A/P_R is high. The gain from multistage filters is realized only for certain values of P_A/P_R as shown by Fig. 24. The increase in P_A/P_R gained by splitting up the same amount of L and C into two or three stages is indicated in Fig. 24, the inductor and capacitor having the same "lumped" values. The attenuation in multistaging is enormous for high X_L/X_C. For lower ratios there may be a loss instead of a gain, as shown by the intersection of the two upper curves.

FIG. 24. Ripple attenuation as a function of filter and ratio of filter reactances.

Other factors may influence the number of filter stages. In some applications modulation or keying may require that a definite size of filter capacitor be used across the load. Usually these conditions result in a single-stage filter, where otherwise more stages might be most economical.

Table 1 shows filter reactors in the negative lead, which may be at either ground or high potential. If low ripple is required in the filtered output, it is usually preferable to locate the filter reactors in the high-voltage lead. Otherwise, there is a ripple-current path through the anode transformer winding capacitance to ground which by-passes the filter reactor. Ripple then has a residual value which cannot be reduced by additional filtering. In a three-phase zigzag full-wave circuit, with center tap used for half-voltage output, separate reactors should be used in the positive leads; placing a common reactor in the negative lead introduces high-amplitude ripple in the high-voltage output.

In rectifiers with low ripple requirements, both filament and anode windings should be accurately center-tapped to avoid low-frequency ripple, which is difficult to filter. Three-phase leg voltages should be balanced for the same reason.

Sometimes an inductor-input filter is tuned as in Fig. 25. The addition of C_1 increases the effective reactance of the inductor to the fundamental ripple frequency. Both regulation and ripple of this type of filter are improved. The filter is not tuned for the ripple harmonics, so the use of high-Q filter inductors is unnecessary. An increase in effectiveness of the filter inductor of about 3:1 can be realized in a single-phase full-wave rectifier circuit. Tuned filters are less effective with three-phase

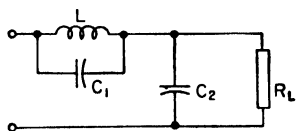



Fig. 25. Shunt-tuned filter.

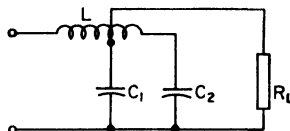



Fig. 26. Series-tuned filter.

rectifiers because slight phase unbalance introduces low-frequency ripple which the filter does not attenuate.

Filters may be tuned as in Fig. 26, where C_1 is connected to a tap near the right end of L, and C_2 is chosen to give series resonance and hence zero reactance across the load at the fundamental ripple frequency. Because of choke losses, the impedance across R_L is not zero, but the resulting ripple across R_L can be made lower than without the use of C_1. Ripple is attenuated more than in the usual inductor-input filter, but regulation is not substantially different.

Where current is small, so that E_{dc} is not unduly affected, filter inductors are often replaced by resistors. This practice is often found in receivers. Plate current for all tubes flows through the first filter-stage inductor. But current for the r-f amplifiers is usually small and flows through resistors in successive filter stages. Such a filter is said to be graded.

15. Capacitor-input Filters.[1] When the filter has no inductor between rectifier and first capacitor, rectifier current is not continuous throughout each cycle and the rectified wave form changes. During the voltage peaks of each cycle, the capacitor charges and draws current from the rectifier. During the rest of the time, no current is drawn from the rectifier, and the capacitor discharges into the load.

Comparison between the rectified voltage of inductor input and capacitor-input filters in a single-phase full-wave rectifier may be seen in Fig. 27. The two tube currents I_1 and I_2 in *a* add to a constant d-c output, whereas in *b* the high-peaked tube currents flow only while the rectified voltage is higher than the average d-c voltage. *Average* current per tube in both cases is half the rectifier output. With large values of capacitance, the rectified voltage in Fig. 27*b* increases to within a few per cent of the peak voltage. Ripple, average rectified voltage output, and rectifier current are dependent on the capacitance, the supply-line frequency, and the load resistance. They are dependent also on rectifier internal resistance because it affects

[1] Waidelich, D. L., Diode Rectifying Circuits with Capacitance Filters, *Trans. AIEE*, **61**, 1161, December, 1941.

the peak value of current which the filter capacitor can draw during the charging interval Δt.

In Figs. 28, 29, and 30, R_S is the rectifier series resistance, including the transformer resistance. Results accurate to within 5 per cent are obtained if the rectifier resistance corresponding to peak current $\hat{I}_P$ is used in finding R_S. The process is cut-and-try, because $\hat{I}_P$ depends on R_S, and vice versa, but two trials usually suffice. Three-phase rectifiers are rarely capacitor-input. The figures are from Lee, *op. cit.*

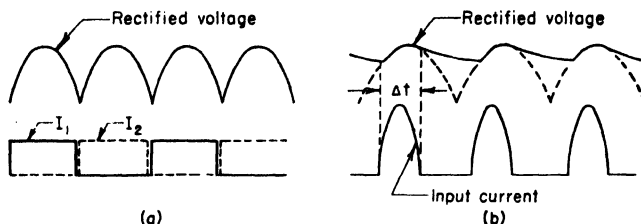



Fig. 27. Voltage and current comparisons in (*a*) reactor-input and (*b*) capacitor-input circuits.

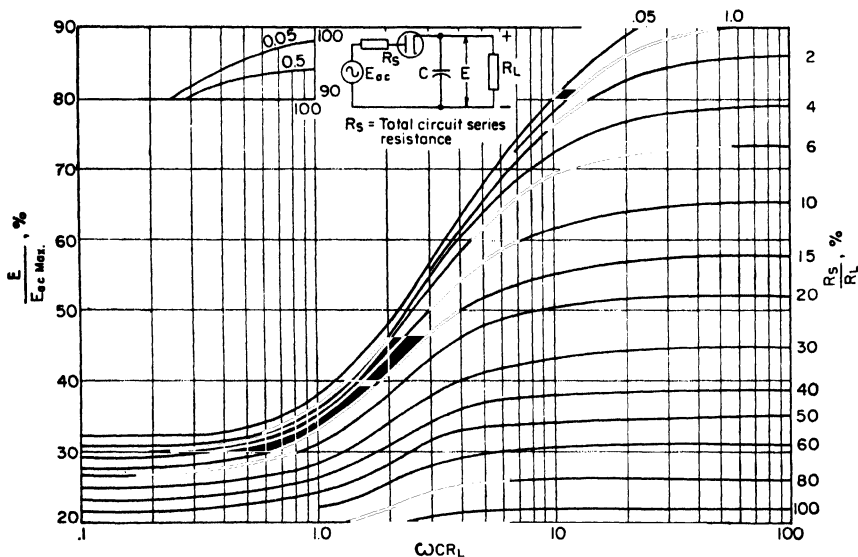



Fig. 28. D-c voltage in half-wave rectifier with capacitor-input filter.

In Fig. 30 the peak current indicates whether the peak current of a given tube is exceeded, and the rms current determines the transformer secondary heating. The v-a ratings are greater, but *ratios* of primary to secondary v-a ratings given in Table 1 hold for capacitor-input transformers also.

One of the assumptions implied at the beginning of Sec. 15, namely, that transformer and rectifier voltage drops are negligibly small, cannot usually be made when capacitor-input filters are used, because of the large peak currents drawn by the capacitor during the charging interval. Such charging currents drawn through finite resistances affect both the d-c output voltage and the ripple in a complicated manner, and simple analysis such as that given for inductor-input filters is no longer possible. Figure 31

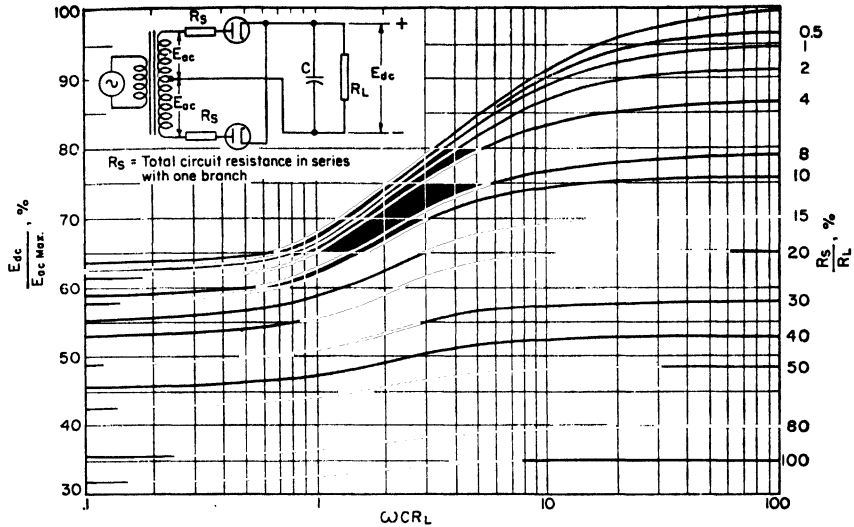



FIG. 29. D-c voltage in full-wave rectifier with capacitor-input filter.

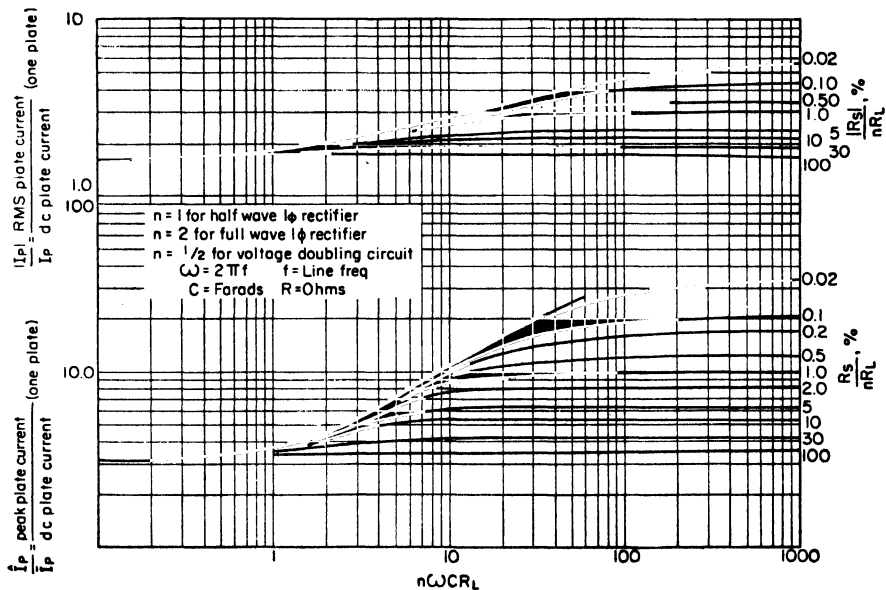



FIG. 30. Peak, average, and rms rectifier current with capacitor-input filter.

is a plot of the ripple in the load of capacitor-input filters with various ratios of source to load resistance, and for three types of single-phase rectifiers. These curves are useful also when resistance is used in place of an inductor at the input of a filter. ω is 2π times the a-c supply frequency, R_L is the load resistance, and R_S the source resistance.

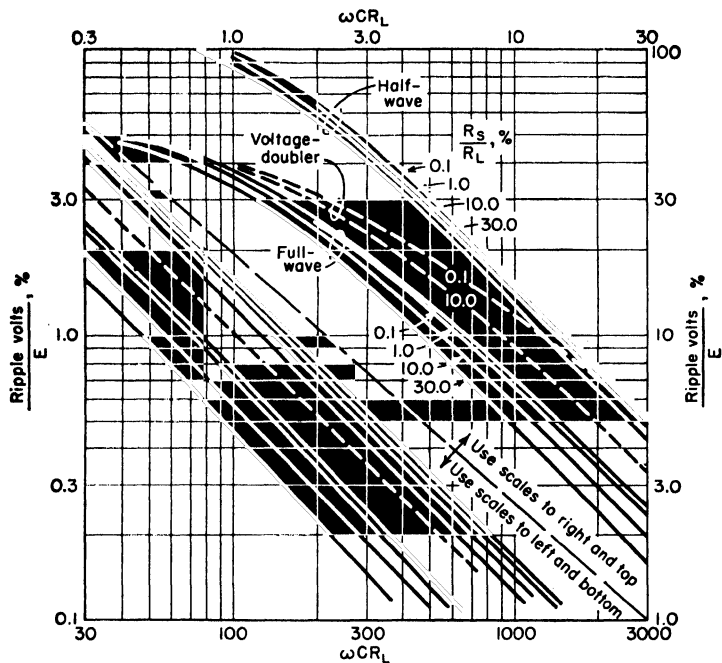



FIG. 31. Rms ripple voltage of capacitor-input filters.

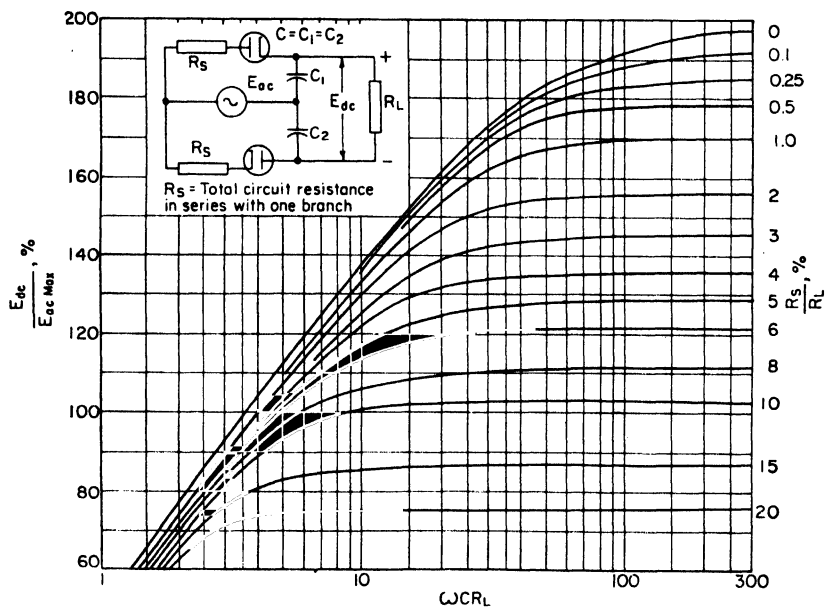



FIG. 32. D-c voltage in full-wave voltage doubler.

When LC filter stages follow a capacitor-input filter, the ripple of the latter is reduced as in Fig. 24, except that the value of P_A must be taken from Fig. 31. When an RC filter stage follows any type of filter, the ripple is reduced in the ratio R/X_c represented by the RC stage.

16. Voltage Multipliers.[1] To obtain more d-c output voltage, the circuit of Fig. 32 is often used. With proper values of circuit elements the output is nearly double the a-c peak voltage. Tube inverse peak voltage is little more than the d-c output voltage, and no d-c unbalance exists in the anode transformer. Current output available from this circuit is less than from the single-phase full-wave circuit for a given rectifier tube. Current relations are given in Fig. 30. The circuit of Fig. 32 is that of a

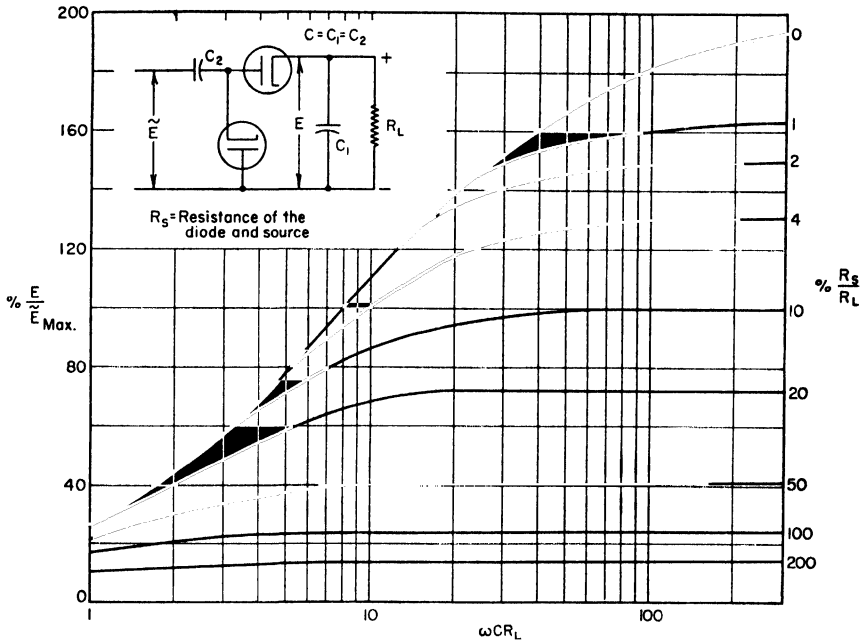



Fig. 33. D-c voltage in half-wave voltage doubler.

full-wave *voltage doubler*. A circuit used in transformerless receivers is the half-wave voltage doubler shown in Fig. 33, together with curves of d-c voltage. Peak currents for this circuit are somewhat less than those of Fig. 30.

Voltage tripling and quadrupling circuits also are used, either to increase the d-c voltage or to avoid the use of a transformer. Such circuits are those of Fig. 34. At $\omega CR_c = 800$, in Fig. 34a[1] E_{dc} is nearly three times and in b nearly four times the peak a-c supply voltage; peak inverse voltage is nearly double the peak supply voltage in both these circuits; ripple is 0.1 per cent in the tripler and 0.3 per cent in the quadrupler; peak current in the tripler is 400 times average, and in the quadrupler 100 times average (see Chap. 15, Sec. 43).

17. Oscilloscope High-voltage Power Supplies. In cathode-ray tubes for oscilloscopes or television receivers, high-voltage low-current sources of d-c power are needed for the accelerating anodes. Voltages range from 1 to 30 kv, and currents are on the order of 1 ma down to 1 μa.

[1] Waidelich, D. L., and H. A. Taskin, Analyses of Voltage Tripling and Quadrupling Circuits, *Proc. IRE*, **33**, 449, July, 1945.

Because of the many turns of wire required in high-voltage 60-cycle transformers, high-frequency power supplies are often used.[1] A tetrode or pentode tube is used in a single- or double-tuned oscillator circuit, with a stepup transformer. The secondary load may be a voltage doubler, tripler, or quadrupler, depending on the voltage needed from the oscillator. Frequencies used in such power supplies range from audio to medium high radio frequency. The oscillator is usually operated class B or C; loaded

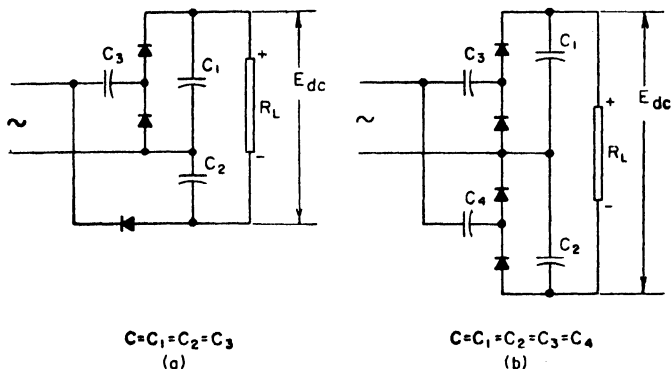



FIG. 34. (a) Voltage-tripling and (b) voltage-quadrupling circuits.

Q ranges from 10 to 20. Lower values of Q result in oscillator instability with load changes.

In television receivers, high voltage is derived from the horizontal deflection circuit, which operates at 15,750 cps. A winding extension on the horizontal output transformer is connected to the half-wave rectifier which supplies high voltage to the cathode-ray tube.

A third alternative is the switching transistor inverter shown in Fig. 35.[2] Here the frequency is fixed by volt-seconds storage in a grain-oriented nickel-steel core. A part of the d-c supply voltage is applied across one half primary which induces base drives of polarity to make the better-conducting transistor conduct still better and make the other transistor a better insulator. This process is regenerative and in a very short time one transistor is driven to saturation while the other transistor is driven to cutoff. Full d-c supply voltage less the small saturated emitter-collector voltage drop is applied across one half primary which induces a voltage constant with time until the core saturates. At this time there is no longer sufficient change in core flux to induce a voltage; so the base drives disappear. The core flux returns from its peak saturated value along a slight slope to its remanent value. This slope starts a voltage reversal which quickly builds up in the opposite direction. Square-wave output of 1,000 to 10,000 cps is produced, the frequency depending on the core material and core area. Toroidal cores are used. Insulation of the high-voltage winding is more difficult than with the usual transformer construction.

FIG. 35. Switching transistor high-voltage supply.

[1] Schade, O. H., Radio-frequency-operated High-voltage Supplies for Cathode-ray Tubes, *Proc. IRE*, **31**, 158, April, 1943.

[2] Chester, M. S., Compact High-voltage Supply Using a Transistor Inverter Circuit, *IRE Convention Record*, Part 6, 1956.

18. Rectifier Regulation. The no-load to full-load voltage regulation[1] of a rectifier comprises three distinct components:

1. The d-c resistance or IR drop
2. The commutation reactance or IX drop
3. The capacitor charging effect

The first component can be reduced to a small value by the use of tubes, transformers, and inductors having low resistance. Mercury-vapor tubes are useful in this respect, as the internal voltage drop is low and almost independent of load-current variations.

Commutation reactance can be kept low by proper transformer design, particularly where the ratio of short-circuit current to normal load is high.

During part of each cycle, both tubes of a single-phase full-wave rectifier are conducting. During this interval one tube loses its current and the other one builds up to normal current. Because of the inevitable reactance in the transformer, this change does not take place immediately. Short-circuit current is initiated which prevents the rectified voltage wave from retaining its normal shape, so that for a portion of each cycle the rectified output is zero.

In a single-phase rectifier, the *average* voltage cut out of the rectified voltage wave by commutation is

$$V_{av} = I_{dc}X/2\pi \tag{6}$$

Commutation reactance drop for polyphase rectifiers is

$$pI_{dc}X'/2\pi \tag{7}$$

where X' = transformer leakage reactance from line to neutral on the secondary side, and p = the number of phases in Fig. 22.

When high winding resistance limits short-circuit current, commutation has less effect than Eq. (6) would indicate. This condition prevails in small rectifiers; the IX drop is negligibly small because of the small transformer dimensions.

In large rectifiers, all rectifier components have low losses to prevent power wastage or overheating, and the IR drop is a very small percentage of the total. At the same time, a large transformer requires careful design to keep the IX drop reasonably small. Therefore, in large rectifiers the IX drop is the dominant cause of regulation. An example with 60 kva rating has 0.7 per cent IR drop and 6 per cent IX drop.

In medium-sized rectifiers the IR and IX drops may be equal, or at least comparable in value. In such rectifiers these two components of regulation do not add arithmetically. Commutation interval θ depends on the short-circuited reactance when resistance is negligible, but if resistance is appreciable θ is related to the ratio X/R exponentially.[2] The increase in regulation caused by commutation reactance may be found from Fig. 36, in terms of d-c output voltage E_{dc}. In this figure the regulation of three widely used rectifiers (single-phase full-wave, three-phase half-wave, and three-phase full-wave) is given

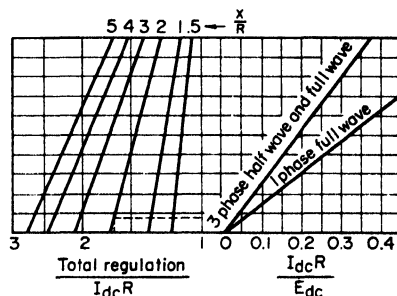



Fig. 36. Increase in rectifier regulation due to transformer reactance.

[1] The term "regulation" is taken to be the loss in voltage within the device divided by the voltage at the output terminals when drawing a given current. Thus

$$\text{Voltage regulation} = \frac{E_{\text{no load}} - E_{\text{specified load}}}{E_{\text{specified load}}}$$

[2] Prince, D. C., and P. B. Vogdes, "Mercury Arc Rectifiers and Their Circuits," p. 216, McGraw-Hill, 1927. See also *Tele-Tech and Electronic Industries*, October, 1953.

in a manner which enables one to proceed directly from the IR component of regulation to total regulation.

X and R are ohms per phase except that X/R ratio is for the whole secondary in single-phase full-wave rectifiers. R in X/R ratio includes primary R in all cases. R in $I_{dc}R/E_{dc}$ is for two windings in three-phase full-wave rectifiers. To obtain total regulation, project $I_{dc}R/E_{dc}$ vertically to one-phase or three-phase line. Project this point to the left to proper X/R line. Abscissa at left gives total regulation. An example is indicated by the dotted line. In this example, the rectifier is three-phase full-wave.

$$E_{dc} = 2{,}000 \text{ volts} \qquad \frac{X}{R} = 2$$

$$I_{dc} = 1 \text{ amp}$$

$$R = 60 \text{ ohms} \qquad \frac{I_{dc}R}{E_{dc}} = \frac{60}{2{,}000} = 3 \text{ per cent}$$

$$X = 120 \text{ ohms}$$

Total regulation = 1.68 × 3 = 5.04 per cent.

If the rectifier had no filter capacitor, the rectifier would deliver the average value of the rectified voltage wave, less regulation components 1 and 2. But with a filter capacitor, there is a tendency at light loads for the capacitor to charge up to the peak value of the rectified wave. At zero load, this amounts to 1.57 times the average value, or a possible regulation of 57 per cent in addition to the IR and IX components, for single-phase full-wave rectifiers. This effect is smaller in magnitude for polyphase rectifiers, although it is present in all rectifiers to some extent.

Suppose that the rectifier circuit shown in Fig. 19*a* delivers single-phase full-wave rectifier output as shown in Fig. 19*b* to an inductor-input filter and thence to a variable load. In such a circuit, the filter inductor keeps the capacitor from charging to a value greater than the average E_{dc} of the rectified voltage wave at heavy loads. At light loads the d-c output voltage rises above the average of the rectified wave, as shown by the typical regulation curve of Fig. 37.

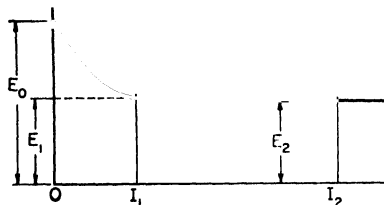



FIG. 37. Rectifier regulation curve.

Starting at zero load, the d-c output voltage E_0 is 1.57 times the average of the rectified wave. As the load increases, the output voltage falls rapidly to E_1 as I_1 is reached. For any load greater than I_1, the regulation is composed of only the two components IR and IX. It is good practice to use a bleeder load I_1 so that the rectifier operates between I_1 and I_2.

Filter elements X_L and X_C determine the load I_1 below which voltage rises rapidly. The filter, if it is effective, attenuates the a-c ripple voltage so that across the load there exists a d-c voltage with a small ripple voltage superposed. A choke-input filter attenuates the harmonic voltages much more than the fundamental, and, since the harmonics are smaller to begin with, the main function of the filter is to take out the fundamental ripple voltage. This has a peak value, according to Fig. 22, of 66.7 per cent of the average rectified d-c voltage for a single-phase full-wave rectifier. Since this ripple is purely a.c. it encounters a-c impedances in its circuit. If we designate the choke impedance as X_L and the capacitor impedance as X_C, both at the fundamental ripple frequency, the impedance to the fundamental component is $X_L - X_C$, the load resistance being negligibly high compared with X_C in an effective filter. The d-c voltage, on the other hand, produces a current limited mainly by the load resistance, provided that the choke IR drop is small.

To prevent capacitor effect the choke must be large enough so that I_{ac} is equal to or less than the bleeder current I_1. This consideration leads directly to the value of choke inductance. The bleeder current I_1 is E_1/R_1, where R_1 is the value of bleeder resistance. The ripple current is the fundamental ripple voltage divided by the ripple circuit impedance.

$$I_{ac} = \frac{0.667E_1}{X_L - X_C}$$

Equating I_1 and I_{ac}, for a single-phase full-wave rectifier

$$R_1 = \frac{X_L - X_C}{0.667} \tag{8}$$

Thus, the value of capacitance also has an effect, but it is minor relative to that of the choke. In a well-designed filter, the choke reactance X_L is high compared with X_C. Therefore, the predominant element in fixing the value of R_1 (and of I_1) is the filter reactor.

Polyphase rectifiers have similar effects, but the rise in voltage is not so great because of the smaller difference between peak and average d-c output. The bleeder resistance for eliminating capacitor effect can be found in general from

$$R_1 = \frac{X_L - X_C}{P_1} \tag{9}$$

where P_1 is the fundamental ripple peak amplitude from Fig. 22, and X_L and X_C are the filter reactances at fundamental ripple frequency.

FIG. 38. Voltage rise in single-phase full-wave rectifier at light loads.

Between load I_1 and zero load, the rate of voltage rise depends upon the filter. Figure 38 shows the voltage rise as a function of the ratio $(X_L - X_C)/R_L$ for a single-phase full-wave rectifier with $IR + IX$ regulation of 5 per cent. Reactances X_L and X_C are computed for the fundamental ripple frequency.

Capacitor-input filters have the voltage-regulation curves shown in Figs. 28, 29, and 32 for their respective circuits. At light loads these filters may give reasonably good regulation, but it is possible to get very poor regulation at heavier loads.

19. Rectifier Transients. Transients which occur when power is first applied are spasmodic. Power may be applied at any instant of a-c voltage cycle, and the suddenly impressed rectifier voltage ranges from zero to E_{pk}. Starting transients are difficult to observe on an oscilloscope because of their random character. It is necessary to start the rectifier several times for one observation of maximum amplitude, and the trace is faint because it appears for a very brief time.

Excessive current inrush, which occurs when a power transformer is connected to a supply line, plagues rectifier design. The phenomenon is associated with core saturation. Heavy initial magnetizing current is drawn from the line, limited only by primary winding resistance and leakage inductance. This heavy current has a peaked wave form which may induce momentary high voltages by internal resonance in the secondary coils and damage the rectifier tubes. Or it may trip a-c overload relays. The problem is especially acute in large transformers with low regulation. A common remedy is to start the rectifier with external resistors in the primary circuit and short-circuit them a few cycles later. Some rectifiers are equipped with voltage regulators which reduce the primary voltage to a low value before restarting.

In some applications the load is varied or removed periodically. Examples of this are keyed or modulated amplifiers. Transients occur when the load is applied (key down) or removed (key up), causing, respectively, a momentary dip or rise in plate voltage. If the load is a device which transmits intelligence, the variation in filter output voltage produced by these transients results in the following undesirable effects:

1. Modulation of the transmitted signal
2. Frequency variation in oscillators, if they are connected to the same plate supply
3. Greater tendency for key clicks, especially if the transient initial dip is sharp
4. Loss of signal power

Although the tendency for key clicks in the signal may be reduced by attention to the d-c-supply filter elements, the clicks may not be entirely eliminated. Where key-click elimination is necessary, some sort of key-click filter is used.

In a choke-input filter, voltage surges are developed across the choke under the following conditions:

1. *Ripple Voltage.* With large rectifier commutation angles, or with grid-controlled rectifiers, a surge occurs once each ripple cycle. In the limit, this surge equals the rectifier peak voltage.

2. *Initial Starting Surge.* This surge adds to output d-c voltage. Under the worst conditions it raises the voltage at this point to twice normal and occurs every time rectifier plate voltage is applied.

3. *Keying or Modulation Transient.* Surge value depends upon L, C, and R_L, and is limited by wave-shape considerations. This occurs each time the key is opened or closed, or load is varied.

4. *Short-circuit Surge.* If R_L is suddenly short-circuited, full d-c voltage appears across the filter reactor until the circuit breaker opens. This occurs occasionally. Rectifiers are sometimes arranged so that, if the short circuit persists, the circuit breaker recloses three times and then remains open.

5. *Interruption of Reactor Current.* This surge voltage is limited only by losses and capacitance of the circuit, and it may be large. Unless the reactor is designed to produce this voltage, it occurs only through accident.

Conceivably, surges 1, 2, and 3 may occur simultaneously and add arithmetically. A reactor insulated to withstand surges 1 plus 2 plus 3 also would withstand surge 4. A reasonable value of peak surge voltage comprising these factors is $2\frac{1}{2}$ times the full d-c working voltage. If surges 1 and 5 are too much for reasonable insulation, the reactor is protected by a gap or other means.

If a rectifier is disconnected from the supply line while the load is off, interruption of plate transformer peak magnetizing current may cause high voltages to appear at random in the windings in much the same way as reactor current interruption causes high voltages. This is especially true if the transformer operates at high core induction. The effect is partly mitigated by the arc energy incident to the opening of the disconnecting switch. But unless the plate transformer is insulated specifically to prevent dangerously high voltages, protective elements may have to be added.

Insufficient attention sometimes is given to the manner in which power-supply lines are brought into buildings. This is particularly important where a rectifier is supplied by overhead high-voltage lines. Because of their relatively high surge impedance, lightning and switching surges occurring on such lines may cause abnormally high voltages in a rectifier and break down the insulation of transformers or other component parts. The likelihood of such surges occurring should be taken into account before the transformers are designed.

Underground-cable power lines impose much less severe hazards because they are protected from lightning strokes, and because they have about one-tenth the impedance of overhead lines. Surges on these cables have much lower values compared with those on overhead lines carrying the same rated voltage. Protection against these surges varies with the type of installation.

The best protection is provided by an indoor power system with an underground cable connecting it to the rectifier. Good protection is afforded by oil-insulated outdoor surgeproof distribution transformers, stepping down to the rectifier a-c power-supply voltage, with an underground cable between the distribution transformer and rectifier. No protection is provided when overhead lines come directly into the rectifier building.

With the trend to dry-type insulation, it is desirable to use lightning arresters on overhead lines where they enter the building. Because of their low impulse ratio, dry-type transformers require additional arresters inside the building. When a line surge is discharged by a lightning arrester, there is no power interruption.

20. Rectifier Filter Charts. Various properties of filters, such as ripple, regulation, and transients, may impose conflicting conditions on rectifier design. To save time in

what otherwise would be a laborious cut-and-try process, charts are used. In Fig. 39 the more usual filter properties are presented on a single chart to assist in arriving at the best filter directly. This chart primarily satisfies ripple and regulation Eqs. (2) and (9) for a choke input filter.[1]

Rectifier connection	50	60	400
3φ half wave	31	37.5	250
3φ full wave	62.5	75	500

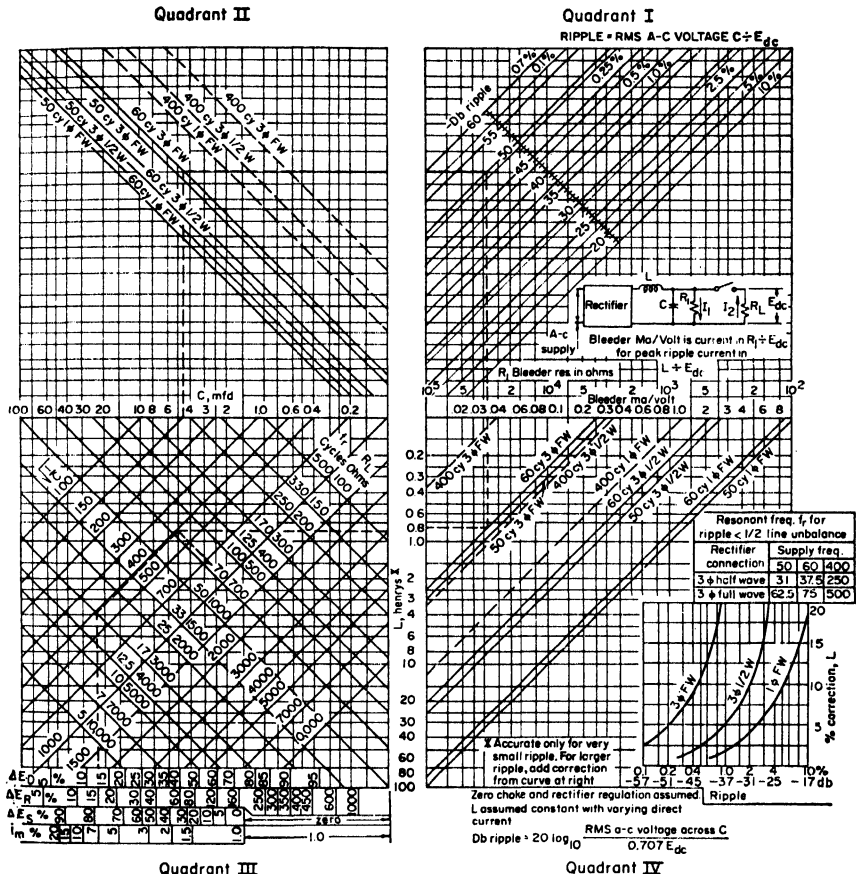



Fig. 39. Choke-input filter chart.

Abscissa values of the right-hand scale are bleeder conductance in milliamperes per volt, and of the left-hand scale, filter capacitance in microfarads. Ordinates of the lower vertical scale are inductance in henrys. Lines representing various amounts of ripple in the load are plotted in quadrant I, labeled in both decibels and rms per cent ripple. In quadrant II, lines are drawn representing different types of rectifiers and supply-line frequencies. A similar set of lines is shown in quadrant IV.

Two orthogonal sets of lines are drawn in quadrant III. Those sloping downward to the right represent resonant frequency of the filter L and C, and also load resistance R_L. The other set of lines is labeled $\sqrt{L/C}$, which may be regarded as the filter impedance. The transient properties of the filter are dependent upon the ratio of $\sqrt{L/C}$ to R_L.

[1] Lee, Reuben, Choke Input Filter Chart, *Electronics*, September, 1949, p. 112.

The L scale requires a correction to compensate for the fact that ripple is not exactly a linear function of L but rather of $X_L - X_C$. The curves in the lower part of quadrant IV give the amount of correction to be added when the correction is greater than 1 per cent.

The following instructions apply for using the chart.

1. Assume suitable value of bleeder resistance or bleeder current I_1 in milliamperes per volt of E_{dc}. This is also steady-state peak ripple current in milliamperes.
2. Trace upward on assumed bleeder ordinate to intersect desired value of load ripple, and from here trace horizontally to the left to diagonal line for rectifier and supply frequency used. Directly under, read value of C.
3. Trace downward on same assumed bleeder ordinate to intersect diagonal line below for rectifier and supply frequency, and read value of L.
4. From desired ripple value, determine correction for L on graph at lower right, and add indicated correction to value of L.
5. Using corrected value of L and next standard value of C, find intersection in quadrant III, and read maximum resonant frequency f_r.
6. Using same values of L and C as in 5, read value of ratio $\sqrt{L/C}$.
7. Under intersection of $\sqrt{L/C}$ with load resistance R_L read values of the four transients illustrated in Fig. 40 (in per cent).

Transient conditions when rectifier is started

Transient conditions with variable load

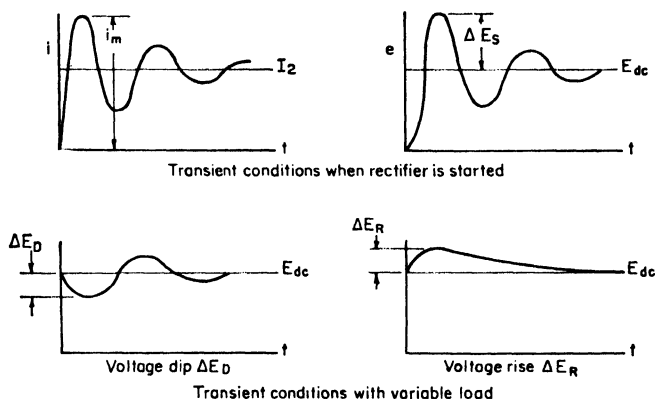



FIG. 40. Four transient conditions in choke-input filter circuit and curves.

Example (shown dashed): Three-phase full-wave 60-cycle rectifier; E_{dc} = 3,000 volts I_2 = 1 amp; I_1 = 96 ma; load ripple = −50 db; balanced line.

Solution:

Bleeder ma/volt = 0.032
C = 4.5 μf (use 5 μf)
Scale value of L = 0.78 henry; corrected value = 0.82 henry
Resonant frequency = 75 cps
Load resistance R_L = 3,000 ohms
$i_m = 7I_2$ = 7 amp; ΔE_D = 12 per cent; ΔE_R = 15 per cent; ΔE_s = 80 per cent

In polyphase rectifiers the possibility exists of enough phase unbalance to impress a voltage on the filter having a frequency lower than the normal fundamental ripple frequency. If the filter L and C resonate near the unbalance frequency, then excessive ripple may be expected. L and C should have a resonant frequency lower than the unbalance frequency to avoid this trouble. Quadrant III of the chart has a series of lines labeled f_r, and the intersection of L and C thereon indicates this resonant frequency. It should be no higher than the value given in the small table on the chart if excessive ripple is to be avoided. This table is based on 2 per cent maximum unbalance in the phase voltages.

For most practical rectifier filters, transient conditions fall within the left-hand portion of the third quadrant. The other conditions sometimes help in the solution of problems in which L and C are incidental, *e.g.*, the leakage inductance and distributed capacitance of a plate transformer.

Although the chart (Fig. 39) applies directly to single-stage untuned filters with constant choke inductance, it can be used for other types with modifications:

1. *Shunt-tuned Choke per Fig. 25.* Figure 39 can be used directly for C, but, for a given amount of ripple, divide the chart values of L by 3 to obtain the actual henrys needed in the choke.
2. *Swinging Choke.* If at light load the filter choke swings to S times the full-load value of henrys, multiply the capacitance obtained from the chart by the ratio S to find the capacitance needed (C_n). The value of L obtained by projecting the bleeder current downward is the maximum or swinging value. It must be divided by S to obtain the full-load value. Transient conditions then may be approximated by using C_n and the full-load value of henrys.
3. *Two-stage Filters.* In a filter with two identical stages, the chart can be used if it is recognized that the ripple is that on the load side of the first choke. For example, if the filter consists of two stages both equal to that in the example given for the single-stage filter, the ripple would not be -100 db but -75 db, because of the fact that the rectifier output has (per Table 1) only 4 per cent ripple, which is -25 db.

The regulation in a two-stage filter, as far as capacitor effect is concerned, depends upon the inductance of the first choke as in the single-stage filter. Therefore the chart applies directly to the inductance and capacitance of one stage. The peak ripple current likewise depends upon the inductance of the first choke, regardless of the location of the bleeder resistor. Transients, however, are more complicated, owing to the fact that the two stages interact under transient conditions.

21. Rectifier Tests. Even though the transformers, chokes, tubes, and capacitors have been tested before assembly of the rectifier, performance tests of the rectifier are desirable. These generally include tests of output, regulation, efficiency, ripple, and input kilovolt-amperes or power factor. Accurate meters should be used, and polyphase rectifiers should have balanced supply voltages. Wiring is tested at some voltage higher than normal, preferably with transformers, tubes, and capacitors disconnected to avoid damage during the test. Ordinary care in testing is sufficient except for regulation tests. If the regulation is low, the difference in meter readings at no load and full load may be inaccurate. Differential measurements are more accurate, such as a voltmeter connected between the rectifier and a fixed source of the same polarity and voltage. Artificial loading of a high-voltage rectifier is often a problem. Water rheostats have been used for this purpose. Load tests, preferably in combination with the transmitter or other apparatus which the rectifier is to supply, are safeguards against field troubles. Operating tests are essential when the load is keyed or modulated, so that overheating or inadequate transformer operation may be detected.

Ripple is measured either with a special hum-measuring instrument or with an RC network arranged to block the d.c. from the measuring circuit. C and R values in the measuring circuit should be so chosen as to avoid influencing the ripple or loading the rectifier transformer. Sometimes capacitance dividers are used for this purpose. The problem of proper values becomes particularly critical with high-voltage low-current rectifiers. The effect of stray capacitance is especially important.

POWER-SUPPLY TRANSFORMERS

22. Construction. Most electronic transformers are small, and for small transformers the shell-type core is usually most suitable because only one coil is required. The magnetic path is divided, half the flux enclosing one side of the coil and half the other. The coil opening is called the window. Between the windows is the core tongue, which is twice as wide as the iron around the rest of the window. The core is built up of thin laminations to reduce eddy-current losses; typical shapes are shown

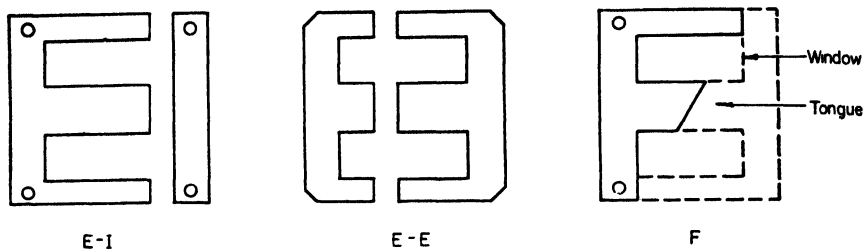



FIG. 41. Shell-type laminations.

in Fig. 41. Alternate stacking of the lamination pairs may be used to reduce magnetic reluctance and keep magnetizing current small. To reduce assembly cost, this alter-

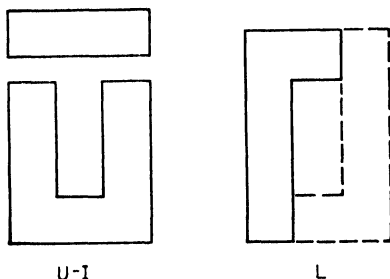



FIG. 42. Core-type laminations.

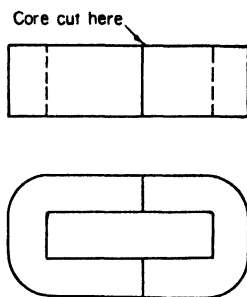



FIG. 43. Type C core.

nate stacking is sometimes done in groups of two or more laminations, with some increase in magnetizing current. A wide range of sizes of shell-type laminations is available. At 60 cycles, common thicknesses are 0.014 in., 0.019 in., and 0.025 in.

Shell-type laminations are made with proportions to suit the transformer. In the E-I shape a scrapless lamination is widely used. Two Es facing each other are first punched, and the punched-out strips are of the right dimensions to form two Is. Then the Es are cut apart. This economy of material is not justified in transformers in which turns per layer, and hence window width, must be reduced relative to window height.

For some applications, the core-type transformer is preferable. In these there is only one magnetic path, but there are two coils, one on each leg of the core. Some core-type laminations are shown in Fig. 42.

Cores wound from continuous steel strip are widely used. One common shape is illustrated in Fig. 43; it is known as the type C core. Steel strip is first wound to the proper build-up on a mandrel. The wound core is then annealed, impregnated with a bond, and cut in two to permit assembly with the coil. After assembly with the coil, the core is held together with a steel band. Several advantages accrue from this construction, which will be discussed in Sec. 23.

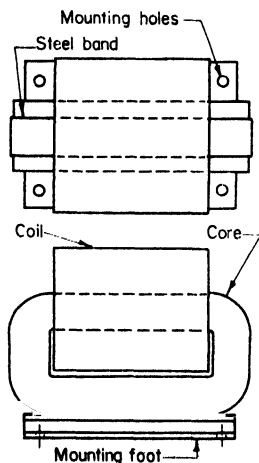



FIG. 44. Single type C core and coil assembly.

Assemblies using two type C cores correspond to shell-type laminations. Because it is simpler to assemble a single-core loop, a single core is often used, especially in

small sizes (see Fig. 44). In 60-cycle service core laminations are usually stacked alternately to produce an overlapping joint. This is approximated in type C cores with ground gap surfaces which fit closely together. Either type of core can be used with core gaps; laminations are stacked butting, with no overlap. The desired amount of gap material, such as fishpaper, is inserted between the gap surfaces.

A shell-type transformer with punched laminations is shown in Fig. 45.

Fig. 45. Shell-type transformer.

23. Materials. In electronic power transformers cores are made chiefly of silicon steel. In usual practice, it is necessary to avoid reaching saturation flux densities, because high exciting currents produce high winding IR drops, high losses, low efficiency, and large size. Curves of induction and core loss are available from manufacturers of laminations. Grades and thickness are designated by numbers such as

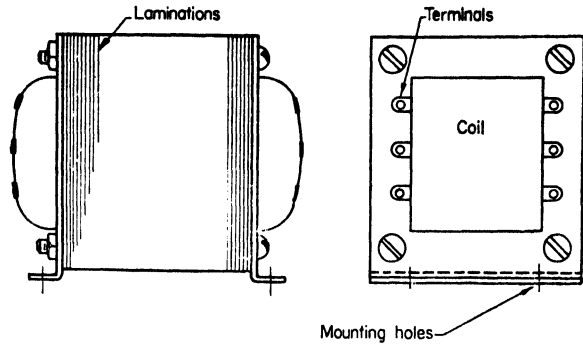
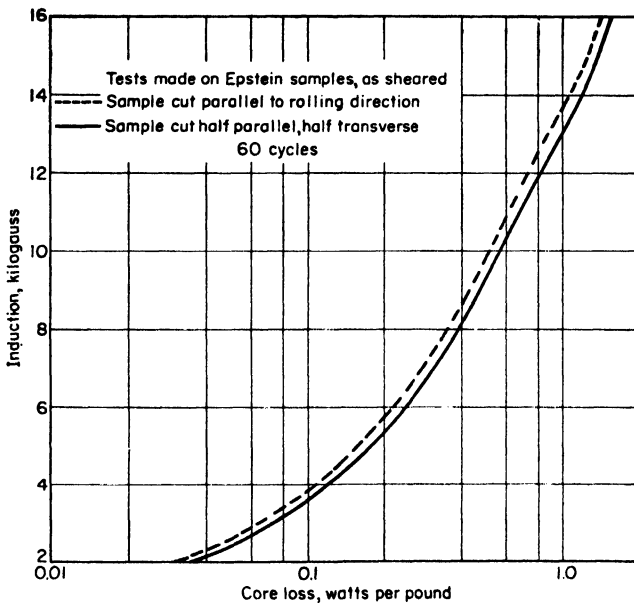



Fig. 46. Core loss at high induction. Armco Trancor M15 grade, 29 gage.

Armco Trancor M15 and Allegheny Transformer A. Hot-rolled silicon-steel laminations are available in 0.014-in., 0.019-in., and 0.025-in. thicknesses, with silicon content of approximately 3 to 4 per cent, and with core losses ranging from 0.6 to 1.2 watts per pound at 10,000 gausses, 60 cps (64,500 lines per square inch). Figures 46 and 47 are

core-loss and exciting v-a/lb for a widely used grade of electronic transformer core steel at 60 cps.

Much work has been done in developing grain-oriented core materials. These materials have a composition similar to that of older nonoriented core material, but grains in the material are oriented by cold rolling. When magnetic flux is established in the lamination, the grains must be aligned in the same direction. If the grains are

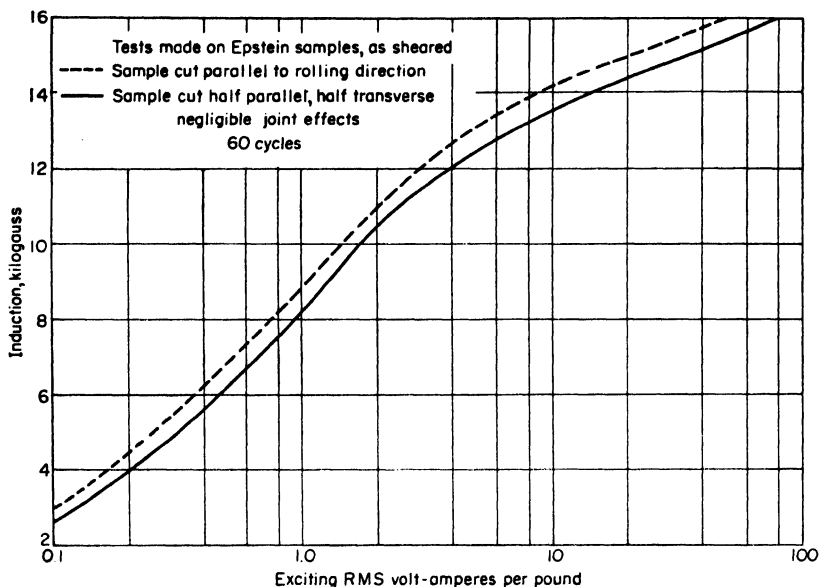



FIG. 47. Exciting rms volt-amperes per pound. Armco Trancor M15 grade, 29 gage.

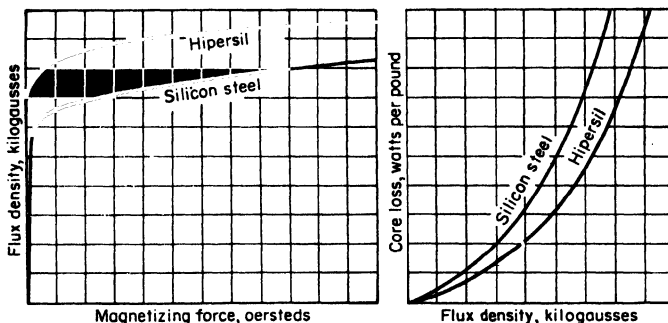



FIG. 48. Induction and core-loss curves of silicon steel and Westinghouse Hipersil at 60 cycles.

already oriented in this direction during the rolling process, much smaller magnetizing force is required to produce the desired flux. Coercive force and hysteresis loss are smaller than in nonoriented steel and permeability is greater.

The development of grain-oriented silicon steel has had a marked effect on transformer size and performance. To illustrate this effect, a comparison is made below between the older nonoriented steel (termed, for simplicity, silicon steel) and Hipersil, a cold-rolled steel in which grain orientation is carried out to a high degree. If core

flux flows in the grain-oriented direction, high core inductions may be realized. Type C cores fulfill this requirement, because the strip is wound in the same direction as the flux path.

The material is rolled in three major thicknesses:

No. 29 gage (about 12 to 14 mils thick) for frequencies up to 400 cps
5 mils thick for frequencies 400 cps and higher
2 mils thick for frequencies in the low and medium r-f bands

Probably the most remarkable property of this material is its high saturation point. In Fig. 48 the comparison is given in terms of a hypothetical 60-cycle working induction using high-grade, conventional silicon steel. If this value is assumed to be 100 per cent, the induction obtained with grain-oriented steel is 130 to 150 per cent, with no increase in magnetizing force. Better space factor can be obtained with strip-wound cores, as is indicated in Table 2.

Table 2. Hipersil Core Data

Hipersil	Thickness, in.	Typical Hipersil space factor, %*	Typical space factor for silicon steel, %*
C-97	0.013	95	90
C-95	0.005	90	80
C-91	0.002	85	70

* Refers to percentage of core volume occupied by metal. The Hipersil figure is for type C cores, and the silicon steel figure is for punched laminations.

Core-loss and exciting v-a/lb for 29-gage Hipersil are plotted in Figs. 49 and 50. Joint reluctance is neglected in Fig. 50.

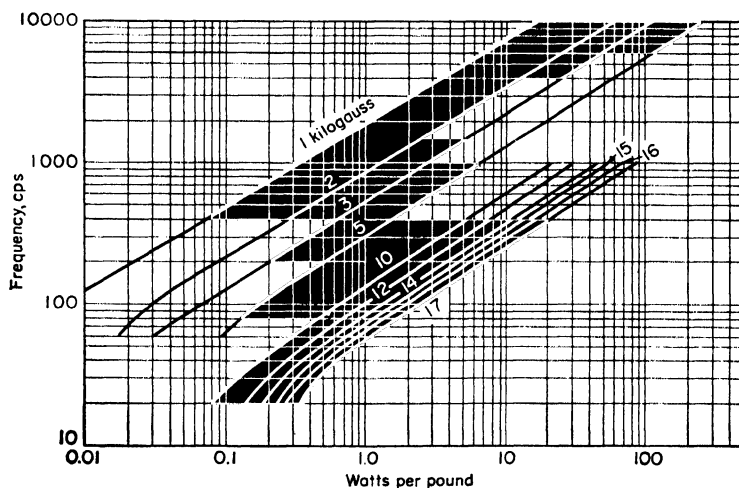



Fig. 49. Core loss in C-97 Hipersil cores (29 gage).

Operating flux density depends upon several factors, such as core material, gage, supply frequency, and operating temperature. For 60 cps, hot-rolled silicon steel laminations may be used at flux densities in the vicinity of 12 kilogausses, and Hipersil in type C cores at about 16 kilogausses. Some manufacturers punch laminations

from grain-oriented sheet; cores of this material may be operated at flux densities intermediate between 12 and 16 kilogausses. The exact value depends on the lamination shape. Lower losses and exciting v-a occur in laminations with long core windows, and with window length in the cold-rolling direction.

At the 400-cycle supply frequency, either 4- or 5-mil Hipersil type C cores offer considerable weight saving when compared with hot-rolled laminations. For the smaller aircraft transformer ratings, it is practicable to operate 4-mil Hipersil at 17 kilogausses and at the lowest supply frequency. Frequency is emphasized because of the relatively wide frequency tolerances encountered in aircraft. With operating frequency lower than nominal, excessive magnetizing current may be drawn by the transformer unless it is designed for this frequency. The result then is overheating or low output voltage.

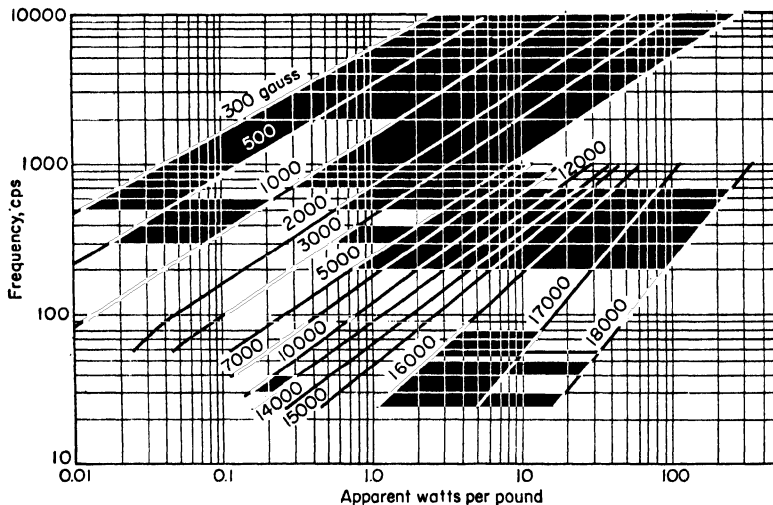



Fig. 50. A-c excitation curve, typical data. C-97 Hipersil cores (29 gage).

24. Windings. Current density in the winding copper is sometimes estimated for design purposes by rules such as 1,000 cir mils per amp. These rules are useful in picking out a first choice of wire size for a given current requirement but should not be regarded as final. Instead, the temperature rise, regulation, or other performance criterion should govern the final choice of wire size. In Fig. 51 the circular mils per ampere are plotted for small enclosed dry-type transformers with Hipersil cores and a winding temperature rise of 55°C; it can be seen to vary appreciably over this range of sizes.

Space occupied by the wire depends on the wire insulation as well as on the copper section. This is especially noticeable in small wire sizes. Table 3 gives the bare and insulation diameters for several common kinds of wire, and Table 4 gives the turns per square inch of winding space. Space usually can be saved by avoiding cotton or silk wire covering and instead using enameled wire with paper layer insulation as in Fig. 52. Thickness of layer paper may be governed by layer voltage; it is good practice to use 50 volts per mil of paper. In coils where layer voltage is low, the paper thickness is determined by the mechanical strength necessary to produce even layers and a tightly wound coil. Table 5 gives the minimum paper thickness based on this consideration.

Space factor may refer to linear spacing, as across a layer, or to the total coil-section area. It is more convenient to use linear space factor in designing layer-wound coils and area space factor in random-wound coils. The values in each case depend largely on the method of winding. For example, it is possible to wind No. 30 enameled wire

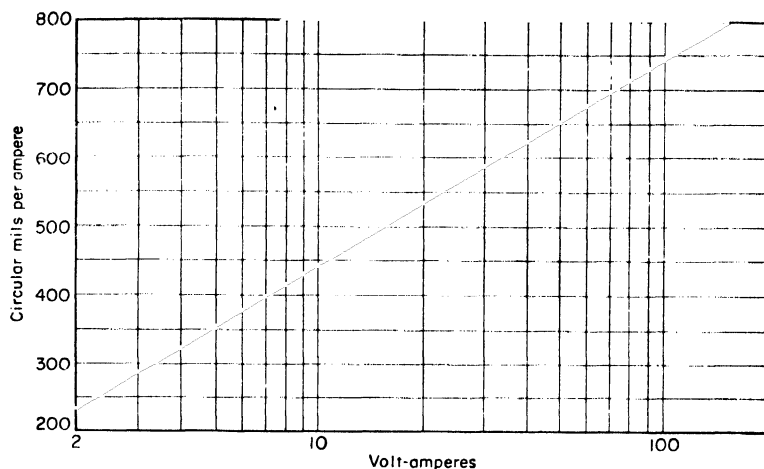



FIG. 51. Wire size in windings of small enclosed 60-cycle transformers.

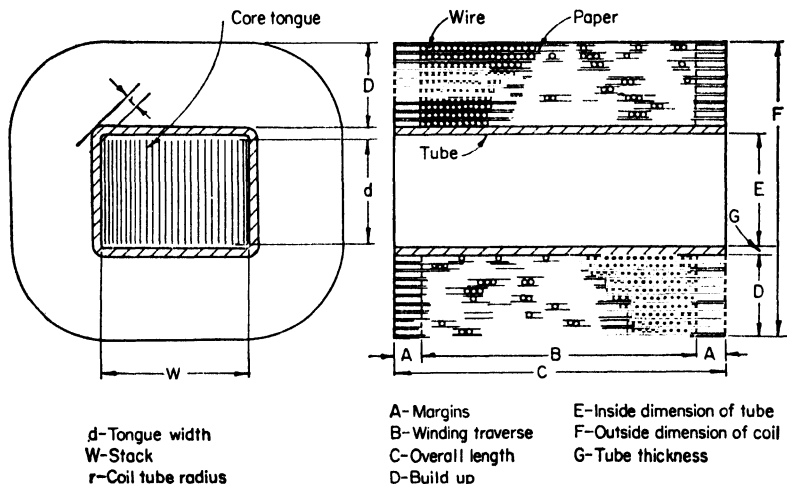



FIG. 52. Paper-insulated coil.

with 97 per cent linear space factor by hand, but with only 89 per cent on an automatic multiple-coil winding machine. Moreover, values of space factor vary from plant to plant. An average for multiple-coil machines is given in Table 5.

Mean length of turn must be calculated for a coil to find its resistance. This may be found by referring to the side view of Fig. 52. Note that there is a small clearance space between core and coil form or tube. Let d be the core tongue and W the stack.

Table 3. Insulated-wire Sizes

B & S gage	Bare diam	Diam of insulated wire								Area, cir mils	Ohms per 1,000 ft at 25°C	Ft per ohm at 25°C	Lb per 1,000 ft
		Single enamel	Double enamel	Single cotton enamel	Single silk enamel	Single cotton	Double cotton	Single silk	Double silk				
44	.0020	.0023								4.00	2,700	.3850	.012
43	.0022	0025								4.84	2,150	.4670	.015
42	.0025	.0029								6.25	1,700	.6050	.019
41	.0028	.0032								7.84	1,350	.7630	.024
40	.0031	.0036	.0039							9.61	1,103	.9550	.030
39	.0035	.0040	.0044							12.25	864	1.204	.038
38	.0040	.0046	.0050							16.00	659	1.519	048
37	.0045	.0051	.0055							20.30	522	1.915	.060
36	.0050	.0057	.0061	.0095	.0075	.0090	.0130	.0070	.0090	25.00	424	2.414	.076
35	.0056	.0064	.0067	.0102	.0082	.0096	.0136	.0076	.0096	31.40	338	3.045	.096
34	.0063	.0072	.0077	.0109	.0089	.0103	.0143	.0083	.0103	39.70	266	3.839	.120
33	.0071	.0080	.0085	.0117	.0097	.0111	.0151	.0091	.0111	50.40	210	4.841	.152
32	.0080	.0090	.0095	.0127	.0107	.0120	.0160	.0100	.0120	64.00	165	6.105	.19
31	.0089	.0100	.0104	.0137	.0117	.0129	.0169	.0109	.0129	79.20	134	7.698	.24
30	.0100	.0111	.0117	.0148	.0128	.0140	.0180	.0120	.0140	100	106	9.707	.31
29	.0113	.0125	.0130	.0162	.0142	.0153	.0193	.0133	.0153	128	83.1	12.24	.38
28	.0126	.0139	.0145	.0175	.0155	.0166	.0206	.0146	.0166	159	66.4	15.43	.48
27	.0142	.0155	.0161	.0192	.0172	.0182	.0222	.0162	.0182	202	52.5	19.46	.61
26	.0159	.0172	.0178	.0210	.0190	.0199	.0239	.0179	.0199	253	41.7	24.54	.77
25	.0179	.0193	.0200	.0234	.0211	.0222	.0262	.0199	.0219	320	33.0	30.95	.97
24	.0201	.0216	.0222	.0256	.0233	.0244	.0284	.0221	.0241	404	26.2	39.02	1.23
23	.0226	.0242	.0247	.0282	.0259	.0269	.0309	.0246	.0266	511	20.7	49.21	1.54
22	.0253	.0271	.0278	.0310	.0287	.0296	.0336	.0273	.0293	645	16.4	62.05	1.95
21	.0285	.0302	.0310	.0344	.0319	.0330	.0370	.0305	.0325	812	13.0	78.25	2.45
20	.0320	.034	.0345	.0385	.0355	.0370	.0410	.0340	.0360	1,020	10.3	98.66	3.09
19	.0359	.038	.0387	.0425	.0395	.0409	.0449	.0379	.0399	1,300	8.14	124.4	3.89
18	.0403	.042	.0431	.0469	.0439	.0453	.0493	.0423	.0443	1,600	6.59	156.9	4.9
17	.0453	.047	.0481	.0521	.0491	.0503	.0543	.0473	.0493	2,030	5.22	197.8	6.2
16	.0508	.053	.0536	.0576	.0546	.0558	.0608	.0528	.0548	2,600	4.07	249.4	7.8
15	.0571	.059	.0605	.0640	.0610	.0621	.0671	.0591	.0611	3,250	3.26	314.5	9.94
14	.0641	.066	.0675	.0711	.0681	.0691	.0741	.0661	.0681	4,100	2.58	396.6	12.4
13	.0719									5,180	2.00	499.3	15.7
12	.0808									6,530	1.59	629.6	19.8
11	.0907									8,235	1.26	794.0	24.9
10	.1019									10,380	1.00	1,001	31.4
9	.1144									13,090	.792	1,262	40.0
8	.1285									16,510	.628	1,592	50.0

Suppose there are several concentric windings. The length of mean turn of a winding V at distance r from the core and having height D is

$$MT = 2W + 2d + 2\pi\left(r + \frac{D}{2}\right)$$
$$= 2(W + d) + \pi(2\Sigma D + D) \qquad (10)$$

where ΣD is the sum of all winding heights and insulation thicknesses between winding V and the core.

Table 4. Turns per Square Inch of Insulated Wire

B & S gage	Single enamel wire	Double enamel wire	Single cotton enamel wire	Single silk enamel wire	Single cotton-covered wire	Double cotton-covered wire	Single silk-covered wire	Double silk-covered wire
42	119,000							
41	96,000							
40	77,000	66,200						
39	62,400	51,800						
38	47,300	40,000						
37	38,400	33,100						
36	30,900	26,900	11,100	17,900	12,350	5,920	20,400	12,350
35	24,500	22,300	9,600	14,900	10,900	5,430	17,200	10,900
34	19,300	16,900	8,430	12,700	9,430	4,900	14,500	9,430
33	15,600	13,900	7,280	10,650	8,130	4,380	12,100	8,130
32	12,350	11,100	6,210	8,740	6,940	3,900	10,000	6,940
31	10,000	9,260	5,330	7,300	5,900	3,510	7,780	5,900
30	8,180	7,300	4,580	6,100	5,100	3,090	6,940	5,100
29	6,430	5,920	3,810	4,950	4,270	2,760	5,670	4,270
28	5,200	4,770	3,280	4,170	3,640	2,360	4,690	3,640
27	4,170	3,880	2,720	3,390	3,030	2,080	3,810	3,030
26	3,330	3,160	2,270	2,780	2,520	1,940	3,120	2,520
25	2,690	2,500	1,820	2,240	2,080	1,460	2,530	2,080
24	2,150	2,030	1,530	1,850	1,690	1,230	2,050	1,720
23	1,710	1,650	1,260	1,490	1,380	1,050	1,650	1,420
22	1,370	1,300	1,045	1,220	1,140	883	1,345	1,160
21	1,100	1,045	846	925	915	729	1,075	943
20	860	850	675	793	730	595	862	836
19	693	668	555	640	597	495	700	628
18	568	540	455	518	490	412	563	510
17	455	432	368	417	395	340	450	412
16	357	350	303	338	320	270	360	335
15	288	273	244	270	260	222	287	268
14	230	220	198	216	210	182	229	222
13	179	176						
12	143	141						
11	114	113						
10	90	90						
9	72	72						
8	57	57						

The mean turn of a winding U just below winding V ordinarily is calculated before that of V. This simplifies the calculation of winding V, the mean turn of which is

$$MT_v = MT_u + \pi(D_u + D_v + 2c) \tag{11}$$

where c is the thickness of insulation between U and V.

Allowance must be made, with many coil leads, for bulging of the coil at the ends and consequent increase of mean turn length.

The placement, insulation, and soldering of leads constitute most important steps in coil manufacture. When coils are wound one at a time, the leads can be placed in the coil while it is being wound. The start lead may be placed on the coil form, suitable insulation may be placed over it, and coil turns may be wound over the insulation. Tap leads can be arranged in the same way. Finish leads must be anchored by means of tape, string, or yarn, because there are no turns of wire to wind over them.

Table 5. Paper-insulated Coil Data
(Courtesy Phelps-Dodge Copper Products Corp.)

B & S gage	Layer insulation, in.	Turns per in.	Space factor, %
44	0.0005	369	85
43	0.0005	340	85
42	0.0005	304	85
41	0.0007	265	85
40	0.0007	239	86
39	0.0007	215	86
38	0.001	193	87
37	0.001	170	87
36	0.001	155	87
35	0.001	140	88
34	0.001	124	88
33	0.0013	110	88
32	0.0013	98	88
31	0.0015	88	88
30	0.0015	80	89
29	0.0015	71	89
28	0.0015	64	89
27	0.0022	57	89
26	0.0022	52	89
25	0.0022	47	90
24	0.0022	42	90
23	0.005	37	90
22	0.005	33	90
21	0.005	30	90
20	0.005	26	90
19	0.007	23	90
18	0.007	21	90
17	0.007	19	90
16	0.010	17	90
15	0.010	15	90
14	0.010	13	90
13	0.010	12	90
12	0.010	10	90
11	0.010	9	90
10	0.010	8	90

In multiple-wound coils, the leads must be attached after the coils are wound. Extra wire on the start turn is pulled out of the coil and run up the side as shown in Fig. 53, with separator insulation between wire extension and coil. Outer insulation covers the wire extension up to the lead joint. A pad of insulation is placed under the joint, and one or more layers of insulation, which insulate and anchor the joint, are wound over the entire coil and the lead insulation. Electrical-grade scotch tape is widely used for anchoring leads. It is important to avoid corrosive adhesives.

Leads should be large enough to introduce only a small amount of voltage drop and should have insulation clearances adequate for the test voltage. These clearances can be found as explained in Sec. 25. In high-voltage transformers it would often be easy to seal the windings if there were no leads; hence lead placement calls for much care and skill. Leads and joints should also be mechanically strong enough to withstand winding, impregnating, and handling stresses without breakage.

25. Insulation. Three classes of insulation are used in dry-type transformers. Class A insulation is organic material such as paper, cotton, silk, varnish, or wire enamel. Class B insulation is mica, asbestos, glass, porcelain, or other inorganic

material with organic binders such as varnish for embedding the insulation. A small amount of other class A material is permissible in a class B coil "for structural reasons," but it should be kept to a minimum.

In general, the vital difference between these classes of insulation is one of operating temperature. Glass-covered wire is preferable to asbestos for space reasons; it is available in approximately the same dimensions as cotton-covered wire. Built-up mica is the usual insulation wrapper material. With special bonds it is flexible enough to wind over coils or layers of wire. Stiff mica plate for lead insulation and mica tubing for coil forms are usually bonded with heat-resistant varnish. Class B insulating material is more expensive than class A and is used only when the advantages outweigh the cost.

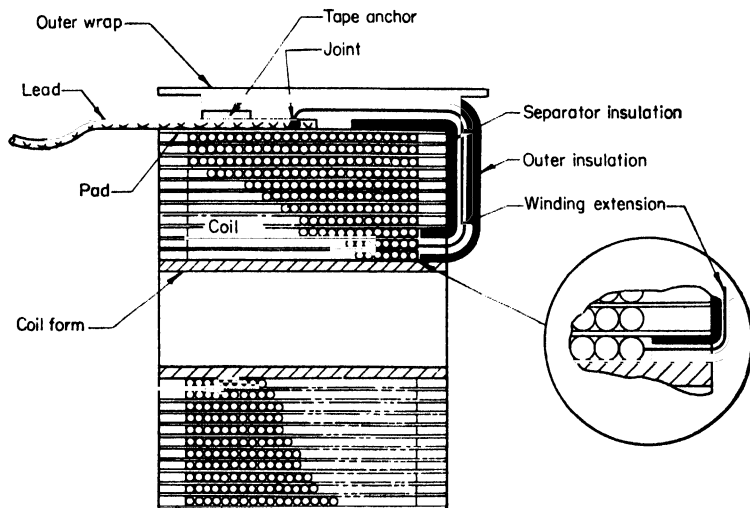



FIG. 53. Start-lead insulation in multiple-wound coils.

A third class of insulation is the *silicones*, which are organic silicates with remarkable thermal and mechanical properties. These materials are coming into use at operating temperatures approaching 200°C. Silicone-treated cloth, silicone rubber, and silicone varnish are already in use. Under development are silicone wire enamel and silicone-bonded mica. They are generally designated as class H insulation.

For apparatus having long service life, AIEE Standard 1 limits the "hottest spot" temperature of impregnated coils as follows:

Class A insulation............	105°C
Class B insulation............	130°C
Class H insulation............	200°C

The usual figure given for dielectric strength is the breakdown value in rms volts at 60 cps in a 1-min test. It is not possible to operate class A insulation anywhere near this value because of the cellular structure of all organic materials. Even after these materials are treated with varnish, many small holes exist throughout a coil structure which ionize and form corona at voltages far below breakdown. With class A insulation (organic materials), the designer must be governed more by resistance of the insulation to corona over a long period than by breakdown strength of the insulation in a 1-min test. For example, a 20-mil thickness of treated cloth will withstand 10,000 volts for 1 min. However, corona starts at 1,250 volts, and operation at any higher voltage would puncture the insulation in a few weeks. It is much wiser to keep a

reasonable margin, say 20 to 30 per cent, below the corona limit than to use a fraction of the 1-min breakdown test. Approximate voltages at which corona is audible are plotted in Fig. 54 as a function of insulation thickness.

Although solid insulation dielectric strength is important, the usual bottleneck for high voltage is creepage distance, such as margins between wire and core along the layers of insulation, or margins between lead joints and frame along the leads and coil sides. A common way of increasing the direct creepage distance across the margins is to use an insulating channel as in Fig. 55*a*. This is especially helpful when the part of the coil adjacent to the core tongue is at low potential and the upper part is at high potential, as in some plate transformers. When the whole coil is at high potential it may be insulated by taping the coil. This is expensive and is avoided wherever creepage safely provides the necessary insulation strength.

Insulation between the start (or finish) turn of the first layer and the core consists of creepage along the margin plus the thickness of the coil form. If, however, the coil lead is brought across the margin and up the side of the coil, the only creepage

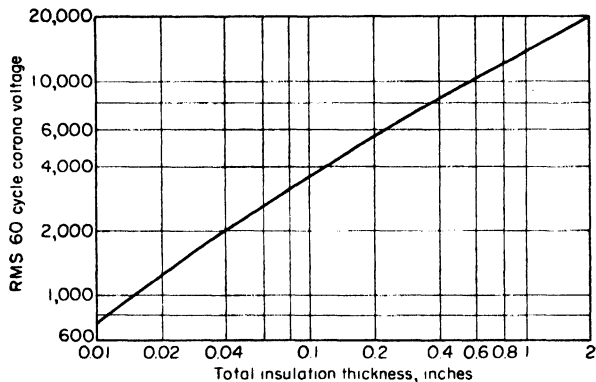



Fig. 54. Corona limit for treated cloth and paper.

distance is the thickness of the coil form. In low-voltage coils this may be enough; in higher-voltage coils, a barrier of insulating material is needed between the coil form and the core, under the spot where the lead is brought out. Such a barrier is provided by outer insulation in Fig. 53. Dimensions of the insulating barrier should be such that a distance at least equal to the coil margin should intervene between the start lead and the core in all directions. The thickness may be the same as the coil form.

In any coil where the finish lead is at the top, there is less difficulty in insulating it. The finish lead has a longer creepage distance to the core if the height of the coil is a greater distance than the margin. It is necessary to avoid using materials on the sides of the coil which would result in any decrease of dielectric strength. In this respect, the creepage strength of some materials with high puncture strength is not good. The last layer of wire may be insulated from the core with a channel as in Fig. 55*a*.

When practical coil margins, even with barriers, are insufficient to support the induced or applied voltage, coils are taped as in Fig. 55*b*. Taping is the most time-consuming but the safest method of insulation.

After a coil is wound it should be impregnated in some sort of insulating liquid which hardens after filling. This is done for several reasons. First, it protects the wire from movement and possible mechanical damage. Second, it prevents the entrance of moisture and foreign matter which might corrode the wire or cause insulation deterioration. Third, it increases the dielectric strength of fibrous insulating materials. Fourth, it assists in heat dissipation from the coil. Single-layer coils may be

dipped in the liquid, drained, and dried, but deeper, thicker coils require the use of vacuum to remove air from the coil and admit the liquid to all parts of the interior. The best mechanical result is obtained when coils are assembled with cores before treatment.

Coils having little or no temperature rise in normal use are impregnated with chemically neutral mineral wax. The wax is melted in a sealed tank and is drawn into another tank in which preheated coils have been placed, and a vacuum is maintained. Coils are removed from the tank, drained, and allowed to cool. Wax treatment provides good dielectric qualities and moisture protection. It is a quick, simple process.

Transformers having operating temperatures of 65°C or higher are impregnated with varnish. Varnish of good grade and close control is essential to achieve thorough filling and dry coils after impregnation. Solventless varnishes, which polymerize to a hard state by baking, are notably useful for the purpose. Filling of the coil may be accomplished by casting the transformer in a mold, or by encapsulation. Encapsulation is readily adapted to irregular coil surfaces and is accomplished by a leakproof

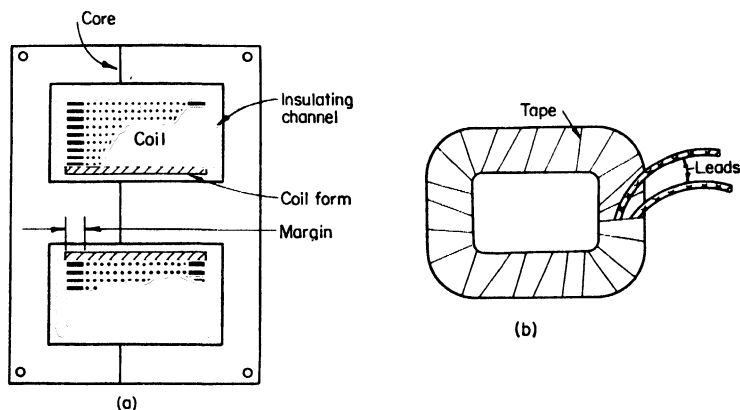



FIG. 55. (*a*) Use of insulation channel. (*b*) Taped coil.

coat before filling. In either process, a good vacuum is necessary to ensure complete filling.

Silicone materials are moisture-resistant. Basic insulation should be inorganic or silicone-treated cloth, tape, laminated sheets, and tubes. Through the use of silicones, some transformers may be designed to have very small dimensions for their ratings. This may be achieved most successfully if the coil insulation comprises only silicone or inorganic materials, including impregnation with silicone varnish. Dielectric strength of silicones is about the same as that of class A materials. Hence the thickness of silicone coil insulation is similar to that for organic materials.

26. Ratings and Color Codes. Core area depends upon voltage, induction, frequency, and turns. For a given frequency and grade of core material, core area depends upon the applied voltage. Window area depends upon coil size or, for a given voltage, upon the current drawn. Since window area and core area determine size, there is a relation between size and v-a rating.

Figure 56 shows the relation between size and rating for small, enclosed, low-voltage, two-winding, 60-cycle transformers having Hipersil cores and class A insulation and operating continuously in a 40°C ambient. The size increases for the same v-a over that in Fig. 56 for any of the following reasons:

High voltage	Silicon-steel cores
High ambient temperature	Low regulation
Lower frequency	More windings

The size decreases for

Higher frequencies	Open-type units
Class B insulation	Intermittent operation

If low-voltage insulation is assumed, two secondary windings reduce the rating of a typical size by 10 per cent; six secondaries by 50 per cent. The decreased rating is due partly to space occupied by insulation and partly to poorer space factor.

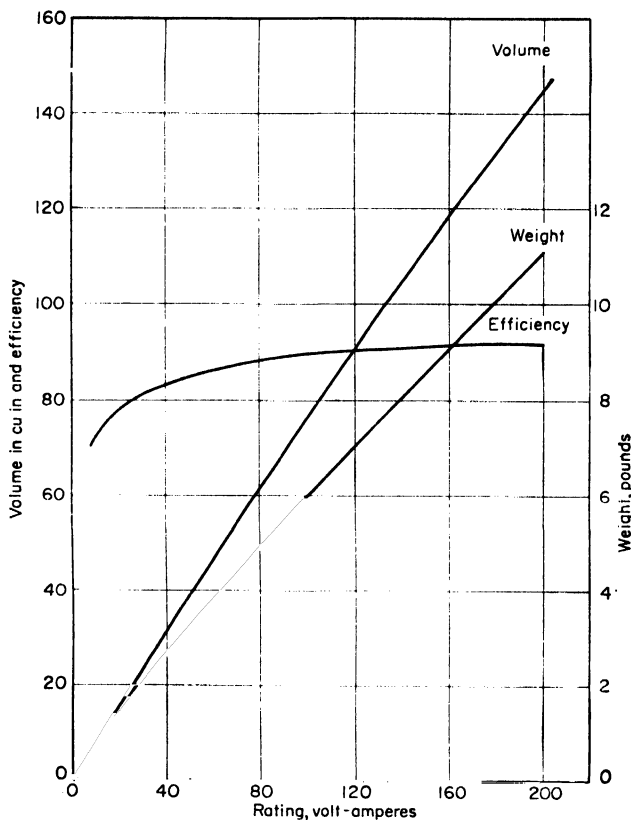



Fig. 56. Size of enclosed 60-cycle transformers.

Transformer Color Code. The leads to transformers of the several types have been standardized as shown below when leads, rather than terminals, are used for external connections.

Power Transformers

Use	Color
Two-wire primary	Both leads black
Tapped primary	Start of winding, black; tap 50/50 black and yellow; finish, 50/50 black and red; if tap is not at center, it is to be nearer the finish end
Rectifier filament	Both leads yellow; tap if used, 50/50 yellow and blue
Rectifier plate	Both leads red; tap 50/50 yellow and red
Amplifier filament winding No. 1	Both leads green; tap if used, 50/50 green and yellow
Amplifier filament winding No. 2	Both leads brown; tap if used, 50/50 brown and yellow
Amplifier filament winding No. 3	Both leads slate; tap if used, 50/50 slate and yellow

I-f Transformers

Use	Color
Plate	Blue
B+	Red
Grid or diode	Green
Grid return or diode return	Black
Secondary tap if used	Green and black

A-f Transformers

Use	Color
Plate	Blue
B+	Red
Other plate for push-pull, start of winding	Blue or brown
Grid or high side of moving coil	Green
Return or low side of moving coil	Black
Other grid for push-pull, start of winding	Green or yellow

27. Simple Transformer Design. It is important that design work be done systematically to save time and to afford a ready means of finding calculations at a later date.

Suppose that a transformer is required to supply filament power for four single-phase full-wave rectifiers having output voltages of 2,000, 500, 250, and 250 volts, respectively, with choke-input filters, as follows:

Primary voltage 100, frequency 60 cycles, four secondaries as follows:

S1:	5 volts	13.5 amp	Insulated for +2,000 volts d.c.
S2:	2.5 volts	10 amp	Insulated for + 500 volts d.c.
S3:	5 volts	3 amp	Insulated for + 250 volts d.c.
S4:	5 volts	2 amp	Insulated for + 250 volts d.c.

Ambient temperature: 40°C

First comes the choice of a core. Data such as those in Fig. 56 are helpful, and so is experience in the modification of such data by the specified requirements. The core selected is a 2-in. stack of laminations described in Fig. 57, and has enough heat dissipation surface for this rating. For silicon steel, an induction of 70,000 lines per square inch is practical. The primary turns

$$N_1 = \frac{E \times 10^8}{4.44 f A_c B} \tag{12}$$

where E = primary voltage, f = line frequency, A_c is the core cross-sectional area, or product of the core tongue width and stack dimension, and B is the core induction. In this transformer, with 90 per cent stacking factor, $A_c = 2 \times 0.9 \times 1.375 = 2.48$ sq in., and the primary turns are found to be 216.

From the sum of the individual v-a figures for windings S_1 to S_4, the transformer rating is found. To it is added an estimate of losses (18 v-a) to obtain the input v-a (135.5) and the primary current (1.36 amp).

Next an estimate of the regulation is made (10 per cent) and added to unity to obtain the multiplier 1.1 in the estimate of secondary turns. From the currents required the wire size for each winding is chosen: P, No. 20; S_2, No. 13; S_3, No. 18; and S_4, No. 20. Round enameled wire is used for each winding except S_1, and for it No. 12 square wire is used to save space. The largest wire is placed next to the coil form to prevent damage in winding to the smaller wires.

Next find out whether the wire chosen will fit in the core window space. Winding height D is calculated for each winding. For each secondary this is the wire diameter, because the wire is wound in a single layer. D for S_1 is slightly larger than the wire dimension to allow for the bulge that occurs when square wire is wound. Twelve turns of S_1 occupy about 1¼ in. of horizontal winding space. The core window is 2¼ in. wide. Subtract ⅛ in. for clearance, leaving 2⅛ in. total coil width. Margins on each side of S_1 are ½(2⅛ − 1¼) = 7⁄16 in. This provides over 8-kv breakdown strength, which is well above the 5-kv test voltage of S_1. Other secondary windings have lower test voltages and wider margins and more than adequate creepage distances.

The 1⁄16-in. Micarta rectangular coil form has a corona voltage of 2,700 rms, which affords about 23 per cent safety factor over the normal operating voltage at the tube fila-

ments. Over S_1 are wound six wraps of 0.010-in.-thick treated cloth, which has 2,600-volt corona limit. S_2 supplies a filament at 500 volts of the same polarity as S_1. Hence only 1,500 volts d.c. or 1,660 volts a.c. occurs across this insulation.

Over S_2 are wound two wraps, over S_3 two wraps, and over S_4 are wound three wraps, all of 0.010-in.-thick treated cloth so that the total cloth thickness amounts to 0.150 in.

The primary winding is wound without layer insulation and with an area space factor of 70 per cent. Cotton is wound in with the wire to form walls 3/16 in. thick on either side of the primary; this accounts for the low space factor and for a 1¾-in. winding traverse. The coil is finished with two layers of treated cloth, a layer of 0.010-in. fishpaper for mechanical protection, and a 0.025-in. serving of untreated cotton yarn or tape to hold it together. The total winding adds up to 0.751 in., leaving 0.124 in. clearance, about the right allowance for winding slack for four secondaries.

Mean turns are figured from Eqs. (10) and (11), with 5 per cent incremental increase of S_2, S_3, and S_4 for leads. With the mean turn values the winding resistances, weights in

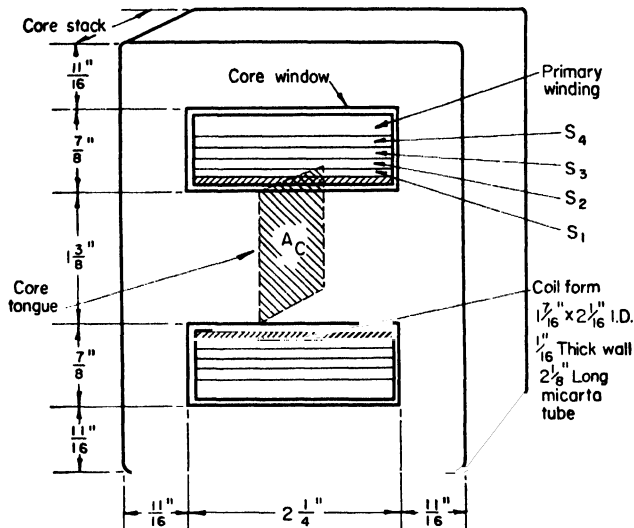



FIG. 57. Dimensions and coil section of filament transformer.

copper, and IR and I^2R for each winding can be found. To S_1, S_2, and S_3 winding resistance is added lead resistance, and the total figure is the sum of the two in each case. Total copper loss is multiplied by 1.3 to correct for 75°C operating temperature. The core weight is 6.8 lb, and the grade of steel used has 1.17 watts per pound at 70,000 lines per square inch. This gives a core loss of 8 watts, and a total of copper and core loss of 20 watts. Added to the input, these losses give efficiency = 85 per cent, which is in accordance with Fig. 56 for this rating. Output voltages on first trial range from 0 to 4 per cent high. S_2 voltage is correct but out of line with the rest. Changing S_2 leads to a larger size makes the per cent voltage drops more nearly alike, and increasing the primary turns to 223 brings all output voltages to correct value within 1.2 per cent. Filament voltage should be kept within 2 per cent for some tubes, to allow for meter error. Primary voltage per layer is equivalent to about 22.7 volts per mil of wire enamel, which is safe practice.

28. Anode Transformer Design. Anode transformers differ from filament transformers in several respects.

1. *Currents* are nonsinusoidal.
2. *Leakage inductance* not only determines output voltage but also affects rectifier regulation in an entirely different manner than with a straight a-c load.
3. *Half-wave rectifiers* carry unbalanced d.c.; this may necessitate less a-c flux density, and larger transformers, than full-wave rectifiers, especially in limited power supplies like aircraft generators. Otherwise, wave-form distortion may result.

4. ***Single-phase full-wave rectifiers with two anodes*** have higher secondary volt-amperes for a given primary v-a rating than a filament transformer.
5. Induced secondary voltage is much higher.

For this last reason, volts per layer are higher in anode transformers, and core windows having proportionately greater height and less width than those in Fig. 57 are often preferable. This trend runs counter to the conditions for low leakage inductance and makes it necessary to interleave the windings. Figure 58 shows the windings of a single-phase full-wave rectifier transformer with the primary interleaved between halves of the secondary. This arrangement is especially adaptable to transformers with grounded center tap. The primary-secondary insulation can be reduced to the amount suitable for primary to ground. This is called *graded* insulation.

Fig. 58. Dimensions and coil section of anode transformer. Construction shown is for shell-type transformer with two Hipersil cores.

Vibrator transformers[1] are similar in design to low-voltage anode transformers, but generally the core induction is lower (8 to 10 kilogausses). Magnetizing current is thus kept small, to aid the vibrator contacts at the instant of switching. This requirement means that vibrator transformers are somewhat larger than anode transformers of the same rating.

In Fig. 58, all flux which follows the core path l_c is mutual flux. Leakage flux is the relatively small flux which threads the secondary winding sections, enters the core, and returns to the other side of the secondaries, without linking the primary. The same is true of flux linking only the primary winding. But it is almost impossible for flux to leave the primary winding, enter the core, and reenter the primary without linking part of the secondary also. The more the primary and secondary windings are interleaved, the less leakage flux there is, up to the limit imposed by flux in the spaces c between sections. These spaces contain leakage flux also; indeed, if there is much interleaving or if the spaces c are large, most of the leakage flux flows in them. Large coil mean-turn length, short winding traverse b, and tall window height a all increase leakage flux.

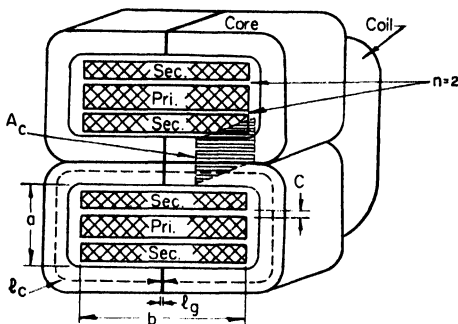
Leakage inductance may be calculated by

$$L_S = \frac{10.6N^2MT(2nc + a)}{10^9n^2b} \tag{13}$$

where L_S = leakage inductance of both windings, henrys, referred to the winding having N turns
MT = mean length of turn for whole coil, in.
n = number of dielectrics between windings (n = 2 in Fig. 58)
c = thickness of dielectric between windings, in.
a = winding height, in.
b = winding traverse, in.

Windings for simple low-voltage 60-cycle transformers may be chosen from winding charts of Fig. 59[2] based upon these conditions:

1. Two untapped concentric windings; primary wound first
2. Operating voltage in both windings less than 1,000 volts
3. Power-supply frequency 60 cycles

[1] EIA Standard RS-177, "Vibrator Power Transformers," February, 1957.
[2] Lee, Reuben, and N. E. Mullinix, Transformer Design Chart, *Trans. IRE*, PGCP, 10, April, 1955.

4. Maximum temperature rise 40°C in 65°C ambient
5. Resistive loads
6. Equal I^2R losses in primary and secondary
7. Solventless-resin-impregnated coils
8. Open-type assemblies like those of Fig. 44
9. Grain-oriented silicon-steel cores

Winding regulation in the two smallest transformers is limited to 15 per cent, and in the next two larger sizes to 10 per cent. In still larger sizes, the 40°C temperature limit held the regulation to less than 10 per cent.

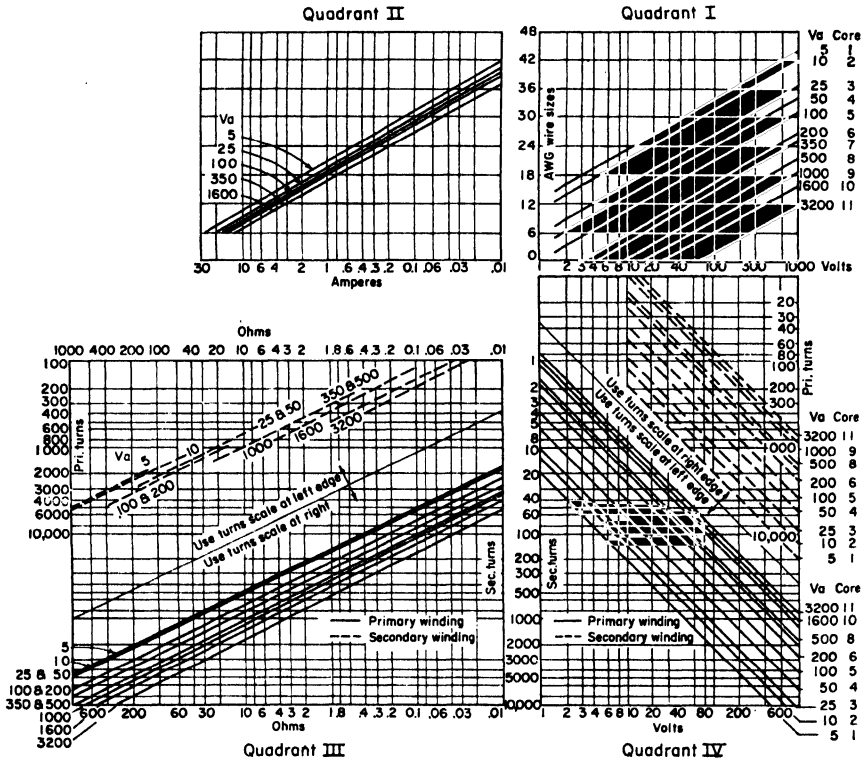



FIG. 59. Low-voltage 60-cycle transformer design chart.

In using the chart, ratings rarely fall exactly on the v-a values assigned to each core. Hence a core is generally chosen with somewhat greater than required rating. Lower regulation and temperature rise than maximum then result. Wire size in quadrant I also increases in discrete sizes, and if the chart indication falls between two sizes the smaller size should be used.

Instructions for Using Fig. 59

1. Choose a core from Table 6 which has a v-a rating equal to or greater than that required.
2. From rated primary and secondary voltages, find number of turns for both windings in quadrant IV.
3. From rated primary and secondary voltages, find wire size for both windings in quadrant I.
4. Project turns across to quadrant III to obtain winding resistances.

Table 6. Transformer Size, Rating, and Regulation

Core	Max v-a rating	% regulation	Total weight, lb	Over-all dimensions, in.
1	5	15	0.38	1¾ × 1¾ × 1¾
2	10	15	0.68	1⅞ × 2⅜ × 1¾
3	25	10	1.2	2¼ × 2⅞ × 2¼
4	50	10	2.2	2½ × 3⅛ × 2½
5	100	8	3.8	3⅛ × 3¾ × 3
6	200	6	6.4	3⅞ × 4¾ × 3⅝
7	350	4	11.0	4⅜ × 5⅜ × 4
8	500	3	15	5⅛ × 6⅛ × 5
9	1,000	2.2	24	5⅞ × 6¾ × 6⅛
10	1,600	1.8	36	7¼ × 8¼ × 7½
11	3,200	1.2	75	9¾ × 12¾ × 8

Departures from the assumed conditions preclude direct application of Fig. 59, but it is still useful as a starting point. For some common modifications, the following notes apply:

1. For each additional secondary winding reduce core maximum rated v-a by 10 per cent. Choose wire size from quadrant II.
2. For 50-cycle transformers, reduce core maximum rated v-a 10 per cent.
3. When permissible temperature rise is higher than 40°C, core maximum v-a equal (v-a in table) $\times \sqrt{\text{temperature rise}/40°\text{C}}$.

Example: A transformer is required for 115/390 volts, 60 cycles, to deliver 77 volt-amperes. This rating falls between the maxima for cores 4 and 5. Using core 5 at 115 volts, read, from Fig. 59, for the primary, 440 turns of No. 22 wire and 3 ohms resistance; for the secondary, 1,700 turns of No. 27 wire and 40 ohms resistance.

29. Power-transformer Tests. A power transformer is tested to discover whether it will perform as required, or whether it will give reliable service life. Some tests perform both functions.

1. *D-c Resistance.* This test is usually made on transformers at the factory as a check on the correctness of wire size.

2. *Turns Ratio.* Once the correct number of turns in each winding is established, correct output voltage can be assured for a coil of given design by measuring the turns. A simple way of doing this is by use of the turns-ratio bridge in Fig. 60. If the turns are correct, the null indicated by the meter occurs at a ratio of resistances

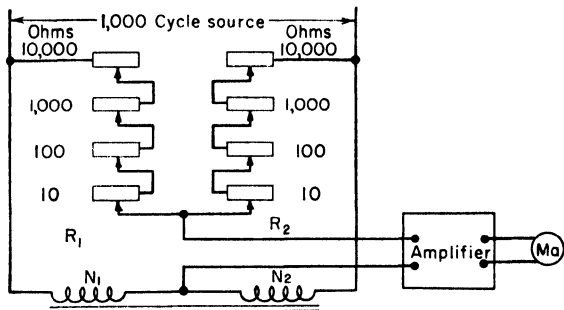



FIG. 60. Turns-ratio bridge.

$$R_1/R_2 = N_1/N_2 \tag{14}$$

If there is an error in the number of turns of one winding, the null occurs at the wrong value R_1/R_2. A filtered source of 1,000 cps is preferable to one of 60 cps for this test.

An accuracy of 0.1 per cent can usually be attained with four-decade resistances. Polarity of winding is also checked by this test, because the bridge will not balance if one winding is reversed.

3. *Open-circuit Inductance (OCL).* There are several ways of measuring inductance. If the Q is high, the check may be made by measuring the current drawn by an appro-

priate winding connected across a source of known voltage and frequency. This method is limited to those cases where the amount of current drawn can be measured. A more general method makes use of an inductance bridge.

4. *Temperature Rise.* Tests to determine whether a transformer overheats are made by measuring the winding resistances before and after a heat run, during which the transformer is loaded up to its rating. Power is applied long enough to allow the transformer temperature to become stable; this is indicated by thermometer readings of core or case temperature taken every half hour until successive readings are the same. Ambient temperature at a nearby location should also be measured throughout the test. The average increase in winding resistance furnishes an indication of the average winding temperature.

5. *Regulation.* It is possible to measure voltage regulation by connecting a voltmeter across the output winding and reading the voltage with load off and on. This method is not accurate because the regulation is usually the difference between two relatively large quantities. Better accuracy can be obtained by multiplying the rated winding currents by the measured winding resistances. If the winding reactance drop is small this works well for resistive loads. To measure reactance voltage drop, a short-circuit test is used. With the secondary short-circuited, sufficient voltage is applied to the primary to cause rated primary current to flow. The quotient E/I is the vector sum of winding resistances and reactances. Reactance is found from

$$X = \sqrt{Z^2 - R^2} \tag{15}$$

where R includes the resistance of both windings and the meter.

Sometimes it is more convenient to measure the leakage inductance with secondary short-circuited on a bridge and multiply by $2\pi f$.

6. *Output Voltage.* Although the method described under (5) above is accurate for two-winding transformers, it is not applicable to multisecondary transformers unless they are tested first with newly calibrated meters to see that all windings deliver proper voltage at full load. Once this is established, values of winding resistance and reactance thereafter can be checked to control the voltage.

7. *Losses.* Often it is possible to reduce the number of time-consuming heat runs by measuring losses. The copper loss is readily calculated by multiplying the measured values of winding resistance (corrected for operating temperature) by the squares of the respective rated currents. Core loss is measured with open secondary by a low-reading wattmeter at rated voltage in the primary circuit. If these losses correspond to the allowable temperature rise, the transformer is safely rated.

8. *Insulation.* If insulation withstands twice normal voltage plus 1,000 volts rms at 60 cps for 1 min, reasonable insulation life is usually obtained. It is possible for a transformer to be extremely underinsulated and still pass this test; conversely, there are conditions under which the rule would be a handicap. Therefore it can only be considered as a rough guide.

The manner of making insulation tests depends upon the transformer. Low-voltage windings can be tested by shorting the terminals and applying the test voltage from each winding to core or case with other windings grounded. Filament transformers with secondaries insulated for high voltage may be tested in similar manner. But a high-voltage plate transformer with grounded center tap requires unnecessary insulation if it is tested by this method. Instead, a nominal voltage of, say, 1,500 volts is applied between the whole winding and ground; after that the center tap is grounded and a voltage is applied across the primary of such value as to test the end terminals at twice normal plus 1,000 volts. Similar test values can be calculated for windings operating at d-c voltages other than zero. Such a test is called an *induced voltage test.* It is performed at higher than normal frequency to avoid saturation. An advantage of induced voltage testing is that it tests the layer insulation.

If insulation tests are repeated one or more times they may destroy the insulation, because insulation breakdown values decrease with time. Successive applications of

test voltage are usually made at either decreased voltage or decreased time. In view of their dubious value, repeated insulation tests are best omitted.

Corona tests are not open to this objection. A voltage 5 per cent higher than normal is applied to the winding, and the leads are run through blocking capacitors to the input of a sensitive radio receiver. EIA standard noise values for this test are given in Table 7.

Table 7. Corona Voltage*

Rms Working Voltage, Kv	Corona Level, μv
Up to 8.6	1,000
8.61 to 15	2,500

* EIA Standard TR-102-B, "Power Transformers for Radio Transmitters."

30. Reactors. Reactors are used in electronic power equipment to smooth out ripple voltage in d-c supplies. Therefore they carry d.c. in the coils. It is common

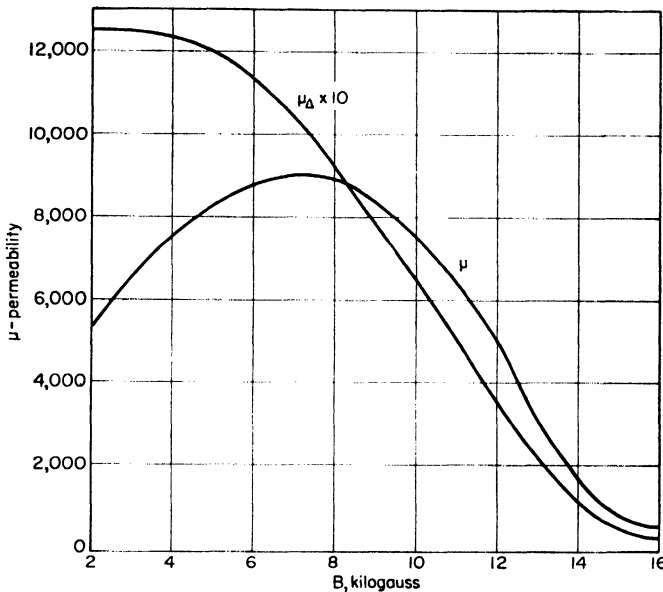



FIG. 61. Normal and incremental permeability of 4 per cent silicon steel.

practice to build such reactors with air gaps in the core to prevent d-c saturation. The air gap, size of core, and number of turns depend upon three interrelated factors: inductance desired; d.c. in the winding; and a-c volts across the winding.

The number of turns, the direct current, and the air gap determine the d-c flux density, whereas the number of turns, the volts, and the core size determine the a-c flux density. If the sum of these two flux densities exceeds saturation value, noise, low inductance, and nonlinearity result. Therefore a reactor must be designed with knowledge of all three of the conditions above.

Magnetic flux through the coil has two component lengths of path: the air gap l_g and the length of the core l_c. The core length l_c is much greater geometrically than l_g, as indicated in Fig. 58, but the two components do not add directly because their permeabilities are different. In the air gap, the permeability is unity, whereas in the core its value depends on the degree of saturation of the iron. The effective length of the magnetic path is $l_g + (l_c/\mu)$, where μ is the permeability for the steady or d-c component of flux.

Reactor design is, to a large extent, the proportioning of values of air gap and magnetic path length divided by permeability. If the air gap is relatively large, the reactor inductance is not much affected by changes in μ; it is then called a *linear* reactor. If the air gap is small, changes in μ due to current or voltage variations cause inductance to vary; then the reactor is nonlinear.

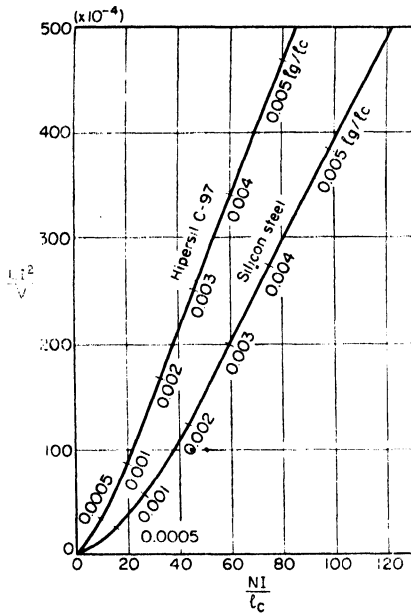



Fig. 62. Reactor energy per unit volume vs. ampere-turns per inch of core.

The curve in Fig. 61 marked μ is the normal permeability of 4 per cent silicon steel for steady values of flux, in other words, for the d-c flux in the core. It is four to twenty times as great as the incremental permeability μ_Δ for a small alternating flux superposed upon the d-c flux. The ratio of μ to μ_Δ gradually increases as d-c flux density increases.

Because of the low value of μ_Δ for minute alternating voltages, the effective length of magnetic path $l_g + (l_c/\mu_\Delta)$ is considerably greater for alternating than for steady flux. But the inductance varies inversely as the length of a-c flux path. If, therefore, the incremental permeability is small enough to make l_c/μ_Δ large compared with l_g, small variations in l_g do not affect the inductance much. For this reason the exact value of the air gap is not important with low alternating voltages.

Reactor size, with a given voltage and ratio of L to R, is proportional to the stored energy LI^2. For the design of reactors carrying d.c., that is, the selection of the right number of turns, air gap, and so on, a simple method was originated by C. R. Hanna.[1] By this method, magnetic data are reduced to curves such as Fig. 62, plotted between LI^2/V and NI/l_c, from which reactors can be designed directly.

The various symbols in the coordinates are:

L = a-c inductance, henrys
I = d.c., amp
V = volume of iron core, cu in.
= $A_c l_c$ (see Fig. 58 for core dimensions)
A_c = cross section of core, sq in.
l_c = length of core, in.
N = number of turns in winding
l_g = air gap, in.

Example: Assume a stack of silicon-steel laminations having a cross section ⅞ by ⅞ in., and with iron filling 92 per cent of the space. The length of the flux path l_c in this core is 7½ in. How many turns of wire and what air gap are necessary to produce 70 henrys with 20 ma d.c. in the winding?

$$A_c = (0.875)^2 \times 0.92 = 0.71 \text{ sq in.}$$
$$V = 0.71 \times 7.5 = 5.3$$
$$\frac{LI^2}{V} = \frac{70 \times 4 \times 10^{-4}}{5.3} = 53 \times 10^{-4}$$

[1] Hanna, C. R., Design of Reactances and Transformers Which Carry Direct Current, *J. AIEE*, **46**, 128, February, 1927.

In Fig. 62 the abscissa corresponding to $LI^2/V = 53 \times 10^{-4}$ is $NI/l_c = 25$ for silicon steel. The ratio of air gap to core length l_g/l_c is between 0.0005 and 0.001.

$$NI/l_c = 25$$
$$N = (25 \times 7.5)/0.020 = 9{,}375 \text{ turns}$$

The total air gap is nearly $0.001 \times 7\frac{1}{2}$ or 7.5 mils; the gap at each joint is half of this value, or 3.75 mils.

The conditions underlying Hanna's method of design are met in most applications. In receivers and amplifiers working at low audio levels, the alternating voltage is low and hence the alternating flux is low compared with the steady flux. Even if the alternating voltage is of the same order as the direct voltage, the alternating flux may be low, especially if a large number of turns are necessary to produce the required inductance; for a given core the alternating flux is inversely proportional to the number of turns. D-c coil resistance is usually fixed by the regulation or size requirements. Heating seldom affects size.

31. Reactors with Large A-c Flux. With higher voltages, it often happens that the a-c flux is no longer small compared with the d-c flux. This occurs in high-impedance circuits where the d.c. is low and the alternating voltage is high. The inductance increases by an amount depending on the values of a-c and d-c fluxes. In a reactor working near the saturation point, increasing a-c flux soon adds to the saturation, which prevents further inductance increase. Saturation of this sort may be avoided by limiting the value of the d-c flux.

Saturation could be reduced by increasing the air gap somewhat. Average permeability increases, and so does the inductance, if the maximum flux density is kept low enough to avoid saturation. Conversely, if saturation is present in a reactor, it is manifested by a decrease in inductance as the d.c. through the winding is increased from zero to full-load value.

In a reactor having high a-c permeability the equivalent length of core l_c/μ is likely to be small compared with the air gap l_g. Hence, it is vitally important to keep the air gap close to its proper value. This is, of course, in marked contrast to reactors not subject to high a-c induction.

It is important that the inductance be the right value. Rectifier regulation can be improved by the use of swinging or tuned reactors. In a swinging reactor, saturation is present at full load; therefore the inductance is lower at full load than at no load. The higher inductance at no load is available for the purpose of improving voltage regulation. The same result is obtained by shunt-tuning the reactor, but here the inductance should be constant from no load to full load to preserve the tuned condition.

In swinging reactors, all or part of the core is purposely allowed to saturate at the higher values of d.c. to obtain high inductance at low values of d.c. They are characterized by smaller gaps, more turns, and larger size than reactors with constant inductance ratings. Sometimes two parallel gaps are used, the smaller of which saturates at full d.c. When the function of the reactor is to control current by means of large inductance changes, no air gap is used. Design of such reactors is discussed in Sec. 33.

The insulation of a reactor depends on the type of rectifier and how it is used in the circuit. Three-phase rectifiers, with their low ripple voltage, do not require the turn and layer insulation of single-phase rectifiers. If the reactor is placed in the ground side of the circuit one terminal requires little or no insulation to ground, but the other terminal may operate at a high voltage to ground. In single-phase rectifiers the peak voltage across the reactor is E_{dc}, so the equivalent rms voltage on the insulation is $0.707E_{dc}$. But for figuring B_{max} the rms voltage is $0.707 \times 0.67E_{dc}$.

32. Linear Reactor Design. A method of design for linear reactors based on three assumptions follows.

1. The air gap is large compared with l_c/μ, μ being the d-c permeability.
2. A-c flux density depends on alternating voltage and frequency.
3. A-c and d-c fluxes can be added or subtracted arithmetically.

Under these conditions, inductance may be found from

$$L = \frac{3.19N^2A_c \times 10^{-8}}{l_g + (l_c/\mu_\Delta)} \quad \text{henrys} \tag{16}$$

where N is number of turns and dimensions are in inches. Term A_c in Eq. (16) is greater than in Eq. (12) because of the space factor of the laminations; if the gap is large A_c is greater still because the flux across it fringes. With large gaps, inductance is nearly independent of μ_Δ. With small gaps, permeability largely controls. There is always a certain amount of gap even with punchings stacked alternately in groups of 1. Table 8 gives the approximate gap equivalent of various degrees of interleaving laminations for magnetic path l_c of 5.5 in.

Table 8. Equivalent Gaps with Interleaved Laminations

0.014-in. Laminations Alternately Stacked	Equivalent Air Gap, In. (Total), with Careful Stacking
In groups of 1	0.0005
In groups of 4	0.001
In groups of 8	0.002
In groups of 12	0.003
In groups of 16	0.004
Butt stacking with zero gap	0.005

In grain-oriented steel cores the air gap may be large compared with l_c/μ_Δ, because of the high *incremental* permeability of these cores. When this is true, variations in μ do not affect the total effective magnetic path length or the inductance to substantial degree. Reactor properties may then be taken from Fig. 63.[1] To keep the reactor linear, it is necessary to limit the flux density. For grain-oriented silicon-steel cores, inductance is usually linear within 10 per cent if the d-c component of flux B_{dc} is limited to 12,000 gausses and the a-c component B_{ac} to 3,000 gausses.

Dotted lines in quadrant I are plots of turns vs. core area for a given wire size and for low-voltage coils, where insulation and margins are governed largely by mechanical considerations. Core numbers in Fig. 63 have the same dimensions and weight as in Table 6.

If the cores increased in each dimension by exactly the same amount, the lines in quadrant I would be straight. In actual cores, several factors cause the lines to be wavy:

1. Ratios of core window height to window width and core area deviate from constancy.
2. Coil margins increase stepwise.
3. Insulation thickness increases stepwise.

Instructions for Using Fig. 63.

1. Estimate core to be used.
2. Divide required inductance by area A_c of estimated core to obtain a value of L/sq in.
3. In quadrant II, locate intersection of L/sq in. and rated I_{dc}.
4. On this intersection, read total gap length l_g and number of turns N. Gap per leg $= l_g/2$.
5. Project intersection horizontally into quadrant I to intersect vertical line which corresponds to estimated core. This second intersection gives d-c resistance and wire size.

Example: Required: 15 henrys at I_{dc} = 50 ma.
Estimate core No. 1.
L/sq in. = 84.3, l_g = 0.015 in., N = 6,000, DCR = 800 ohms.
Wire size = No. 36.
(Example shown starting with small circle.)

A similar chart may be drawn for silicon-steel laminations, but to maintain linearity lower values of flux density should be used.

Equation (16) gives the inductance of a linear reactor with an air gap. It is assumed that 85 per cent of the core flux is confined to the cross section of core face adjoining the gap. The remaining 15 per cent of the core flux "fringes" or leaves the sides of the core, thus shunting the gap. Fringing flux *decreases* the total reluctance of the magnetic path and *increases* the inductance to a value greater than that calculated

[1] Lee, Reuben, Linear Reactor Chart, *Electronics*, March, 1955, p. 208.

from Eq. (16). Fringing flux is a larger percentage of the total for larger gaps. Very large gaps are sometimes broken up into several smaller ones to reduce fringing.

If it is again assumed that the air gap is large compared with l_c/μ, the reluctance of the iron can be neglected in comparison with that of the air gap. For a square stack of punchings, the increase of inductance due to fringing is

$$\frac{L'}{L} = \left(1 + \frac{2l_g}{\sqrt{A_c}} \log_\epsilon \frac{2S}{l_g}\right) \tag{17}$$

where S is the window length plus $l_g/2$.[1]

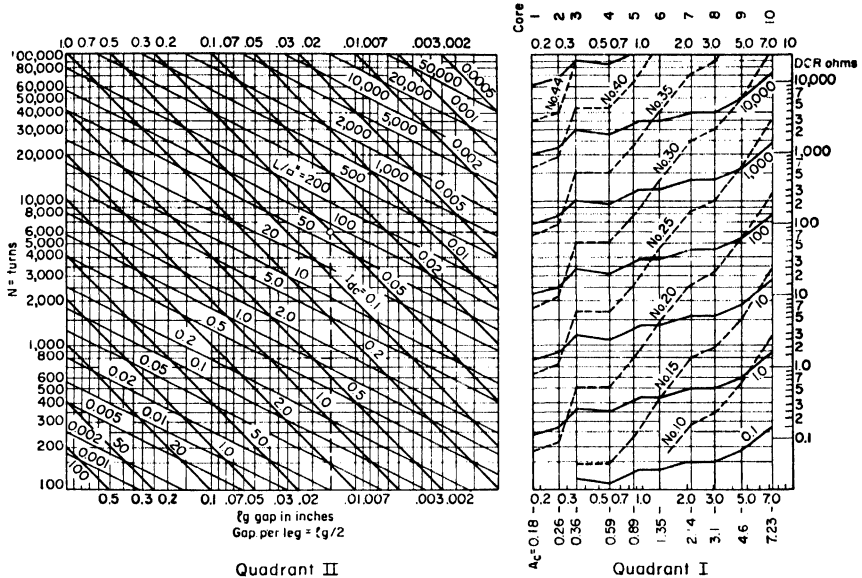



FIG. 63. Linear reactor design chart.

33. Saturable Reactors. The voltage induced in a winding usually far exceeds the resistance drop in that winding. In other words, winding open-circuit reactance usually is much greater than winding d-c resistance. Also a relatively small amount of d.c. flowing into the winding of a transformer, in the core of which there is no air gap, causes the core to saturate. Thus, the reactance of the transformer may be varied by a small amount of d-c power. Now, if one winding of a transformer is connected between an a-c supply and a load, the amount of power delivered to the load may be controlled by a small amount of d-c power flowing in another winding. Because of the fact that open-circuit reactance ordinarily exceeds d-c resistance, the possibility of power amplification is inherent in a transformer. When one winding of a transformer is used for d-c control power and another for a-c output power, the transformer is called a saturable reactor.

To keep the d-c source from being a short circuit to the a-c voltage in the power winding, common practice is to use two reactors, the d-c windings of one of which are reversed, while the a-c windings add normally. Connections of this sort are shown in Fig. 64*a*, with the a-c windings in series; it is possible to connect them in parallel as in Fig. 64*b* to allow more load current to flow at lower a-c voltage.

[1] Partridge, G. F., The Inductance of Iron-cored Coils Having an Air Gap, *Phil. Mag.*, **22**, 675 (7th Series), July-December, 1936.

When there is zero d.c. in the control windings of Fig. 64, both reactor impedances are large and prevent any load current except exciting current from flowing throughout the a-c voltage cycle. When d.c. is applied to the control windings, impedance remains large for the first part of a cycle, until saturation flux density is reached. Then reactor impedance is reduced and a large load current may flow. With grain-oriented core material, the change from high to low impedance is abrupt, and the load current wave form is as shown by i_L in Fig. 65*a*. Only the exciting current flows in the load during the interval 0 to θ_1. Then saturation is reached and load current suddenly rises to a large value. From θ_1 to π, i_L has sinusoidal shape. During the next half cycle, this load-current shape is repeated but in the reverse direction.

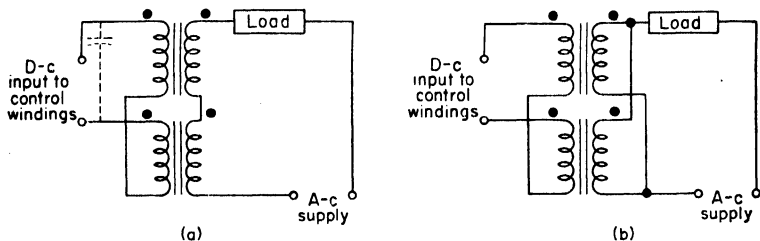



FIG. 64. (*a*) Series-connected and (*b*) parallel-connected saturable reactors.

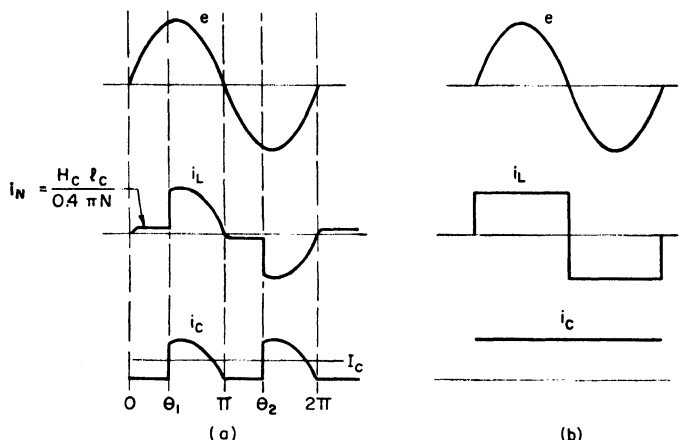



FIG. 65. Saturable reactor currents with (*a*) high and (*b*) low control-circuit resistance.

For a 1:1 turns ratio in each reactor, current i_c in each control winding equals i_L minus the exciting current. In one reactor, because of the reverse connection, current i_c flows in the opposite direction. Total current in the control circuit is as shown by the lowest trace of Fig. 65*a*, the average value of which is the input direct current I_c. Thus load current contains fundamental and odd harmonics, whereas control current contains only even harmonics. If sufficient control current flows to saturate the cores over the full cycle, load current also flows over the full cycle and is sinusoidal in wave form. For turns ratios other than unity, load and control currents are inversely proportional to turns ratio.

In the foregoing it was assumed that control current was free to assume the shape shown in Fig. 65*a*. This is true on a 1:1 turns-ratio basis only if the control circuit impedance is small. If total control circuit resistance is denoted by R_C and load resistance by R_L, for $R_C \ll R_L$, load and control currents are sine waves, or portions

thereof. If the opposite is true, namely, $R_C \gg R_L$, control current wave shape is determined by R_C. For very large R_C, control current is continuous, and the current wave shapes approach those in Fig. 65*b*. In this figure, d-c source impedance is large, even harmonics cannot flow, magnetization is "constrained," load current is flat-topped, and voltage across the reactor is distorted considerably. This distortion can be overcome by the use of a capacitor across the control coils as shown dotted in Fig. 64. When the reactors are parallel-connected as in Fig. 64*b* even harmonics may flow in the load windings, and the capacitor is not needed.

Fig. 66. Windings and core flux paths in a saturable reactor.

Sometimes the two cores are combined into one, in the manner shown in Fig. 66. This is called a three-legged reactor, with one d-c coil and two a-c coils. Figure 66 shows the relative paths for the a-c and d-c fluxes. Equal turns in the a-c coils set up equal a-c magnetomotive forces which cancel in the center leg, and cause flux to flow as indicated by the solid line. No fundamental a-c voltage is induced in the d-c coil, but d-c flux flows in both outer legs as indicated by the dotted lines. A change of current in the d-c coil causes a change in total flux linking the a-c coils and hence a change of inductance. A-c coils may be connected in parallel instead of series, provided that equal turns in each coil and the flux polarity of Fig. 66 are maintained; for the same total number of turns the inductance is halved and the a.c. doubled. The middle core leg shunts the even harmonics of a-c flux.

Fig. 67. Magnetization curves for 4 per cent silicon steel.

It is only in grain-oriented core materials that the wave shapes of Fig. 65*a* are even approximated. In unoriented core steel, wave shape is much more rounded, and control current bears less resemblance to load current. Figures 67 and 68 indicate the contrast in saturation control afforded by unoriented silicon steel and oriented nickel steel. In grain-oriented steel cores there is an approximately linear relationship between d-c ampere-turns per inch and a-c ampere-turns per inch over a large range of

flux density. Moreover, the a-c NI/in. for a given d-c NI/in. changes only slightly with a-c flux density over this range. In Fig. 68, each of the lines for a given value of control magnetizing force is nearly vertical. For a given value of control NI/in., load current is almost independent of flux density and therefore of a-c supply voltage.

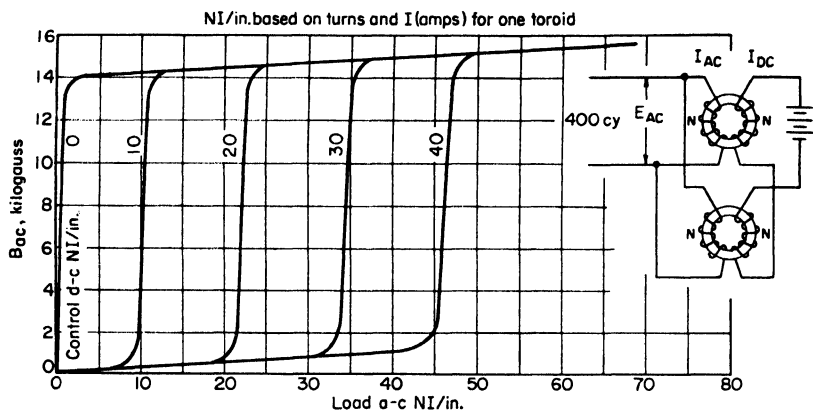



FIG. 68. Typical magnetization curves for 0.002-in. grain-oriented nickel-steel toroidal cores.

Grain-oriented nickel steel cores are wound of thin strip into toroids. The remainder of this section is based on the use of grain-oriented core steel.

If we choose three line voltages corresponding to flux densities within the linear portions of Fig. 68, and plot the d-c control vs. a-c load ampere-turns per inch, the curves of Fig. 69 result. If, instead of NI/in., average load current is plotted, Fig. 69 gives the transfer curves for a simple magnetic amplifier. The curves are symmetrical

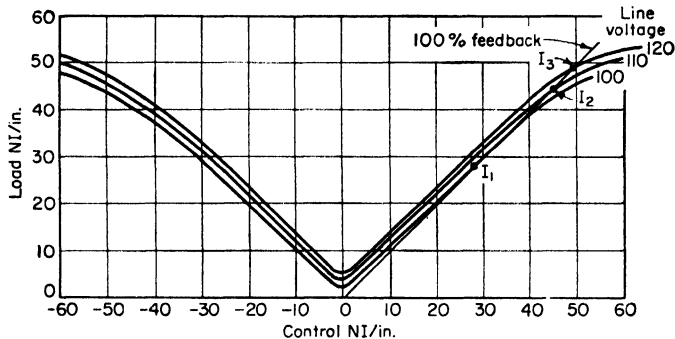



FIG. 69. Simple magnetic amplifier transfer curves with line-voltage variations.

about zero ampere-turns. The difference between the transfer curve and a straight line indicates the degree of nonlinearity in the amplifier for any load current. With grain-oriented core material the a-c load current is nearly independent of supply voltage for a-c inductions less than saturation.

Because of the inductance of the reactor coils, when a change is made in the control winding d.c., load current does not change immediately to its final value. An interval of time, called *response time*, elapses between the change in control current and the establishment of a new steady value of load current. For series-connected reactors,

an average or equivalent control circuit inductance may be found from the relation

$$T_d = \frac{L_c}{R_c} = \frac{R_L}{4fR_c}\left(\frac{N_c}{N_L}\right)^2 \tag{18}$$

where T_d = time for load current increment to reach 63 per cent of final value
L_c = equivalent total control coil inductance, henrys
R_c = total control circuit resistance, ohms
R_L = load resistance
f = line frequency
N_c = turns in control winding
N_L = turns in load winding

For parallel-connected reactors the time of response is appreciably longer. If a rectifier is interposed between the reactor and load, it is possible to obtain sufficient power from the rectifier to supply most of the control power. If the control power from the rectifier furnishes the ampere-turns represented by the straight line in Fig. 69, the amplifier is said to have 100 per cent "feedback." It is then necessary for the control winding to supply only the amount represented by the horizontal difference between the transfer curve and the straight line. This greatly increases the amplification of a pair of reactors; response time increases, but gain per second increases, transfer curve is asymmetrical, and gain is more dependent on supply voltage than in simple magnetic amplifiers. These magnetic amplifiers form a separate subject, for which the reader is referred to the literature, which is extensive.[1]

34. Filter Capacitors. Filament-type rectifiers come up to operating temperature in a few seconds and begin supplying rectified voltage before indirectly heated cathode tubes (load) are warm enough to take their rated currents. Dry-contact rectifiers also operate quickly; the warm-up time is zero. Thus peak voltages are used in specifying filter-capacitor ratings. For example, the first capacitor should be rated to stand 1.41 times the rms transformer secondary voltage; for a 400-0-400-volt secondary the peak is 400 × 1.41 = 564 volts (half the secondary supplied peak voltage in a full-wave rectifier). The second and third filter capacitors also may receive practically full peak voltage when no load current is drawn. For some circuits, an appreciable percentage of the total load current flows in a voltage divider as a waste or "circulating" current. In this case, the second and third capacitor ratings can be proportionally lower by the amount of the voltage drop in the chokes or filter resistors.

While it is true that good filter capacitors will stand voltages greater than their d-c ratings, the regular practice of applying such higher voltages is seldom advisable from a life standpoint. Electrolytics, especially, are subject to rapid deterioration if worked above their ratings.

In general, the paper capacitor has no polarity and extremely low leakage current, changes little with wide temperature variations, is more expensive, weighs more, takes up more space for a given capacitance, and has long life. Paper capacitors are made for practically any high voltage, but seldom rated lower than 400 volts d.c. working.

Electrolytic capacitors are polarized, have appreciable leakage currents, somewhat restricted temperature ranges; single units are available for voltages up to 600 d.c. working, are relatively inexpensive, light in weight, and small in volume for a given capacitance. They have shorter life than paper capacitors. Reasonably sized electrolytics can be made with several thousand microfarads at such low working voltages as 15, 30, and 50.

The life and stability of an electrolytic capacitor are largely determined by the operating temperature of its electrolyte. The temperature depends on how much

[1] For example see Ramey, R. A., On the Control of Magnetic Amplifiers. *Trans. AIEE*, **70**, 2124, 1951. Miles, James G., Bibliography of Magnetic Devices and the Saturable Reactor Art, *Trans. AIEE*, **70**, 2104, 1951 (containing a list of 901 references). Storm, H. F., "Magnetic Amplifiers," Wiley, 1955, contains extensive bibliography.

heat is generated inside the capacitor and the rate at which heat is taken away by the surroundings. Electrolytic capacitors are normally rated for an ambient temperature of 60°C. Specially processed units can operate at 85°C. The low-temperature operating limit is seldom a problem for indoor use, but capacitors for use at temperatures as low as −40°C can be made with the effective rated capacitance as 40 per cent that at normal room temperature.

Typical specifications for electrolytic capacitors include rated capacitance and tolerance, operating frequency, rated voltage, normal operating voltage, peak operating voltage, operating temperature range (max and min), type of circuit used (full-wave, half-wave, doubler, etc.).

Paper filter capacitors are usually oil-filled or wax-impregnated. The selection of the proper oil or wax depends largely on the temperature range to be encountered, either in use or idle. Complete specification for a paper-dielectric capacitor requires: capacitance and tolerance; maximum instantaneous peak voltage; a-c ripple frequency; half-wave or full-wave; average ambient temperature; maximum and minimum temperature expected when not operating; special operating conditions such as high humidity, dusty location, high elevation.

High-voltage capacitors are immersed in oil, askarel, or other special insulating fluids. TV high-voltage dry-type ceramic capacitors rated up to 500 μμf, 30,000 volts, are available with +50 to −20 per cent standard tolerance. By-pass ceramic-disk capacitors are rated up to 0.02 μf at 500 volts and up to 100 μμf at 6,000 volts. Tantalum capacitors are useful for low voltages where maximum ratings of 300 μf at 6 volts or 25 μf at 125 volts are required. Both tantalum and ceramic-disk capacitors have large capacity tolerances and are especially suited to printed circuits and other applications in which small dimensions are necessary.

POWER-SUPPLY-VOLTAGE CONTROL

35. Control and Regulation. Power-supply voltage is varied to obtain suitable volume, brightness, or other output levels. This is accomplished by controlling the transformer primary voltage, or turns ratio, or secondary voltage, by variable autotransformers, tap switches, tubes, or magnetic amplifiers. Once the right voltage is established, it may be necessary to hold it constant in spite of line voltage or load changes. Constant d-c output requires some kind of regulator on most power systems. It is with this meaning of the word "regulation," as distinct from the no-load to full-load regulation with constant line voltage described in Sec. 18, that the following sections are concerned.

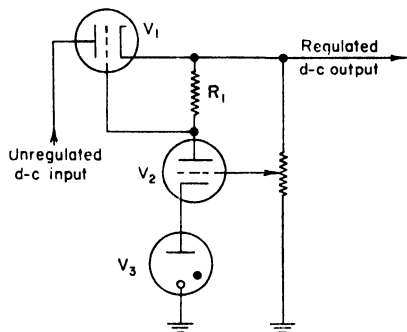



Fig. 70. Vacuum-tube voltage regulator.

36. Tube Regulators. In low-power applications vacuum-tube regulators are used. Output voltage is determined by a V-R cold-cathode tube used as a reference. A control tube is connected between rectifier output and load, with the grid responsive to changes in d-c output voltage. In this way, control-tube plate resistance is varied automatically to maintain constant output voltage.

Figure 70 is the diagram of such a regulator.[1] Here V_1 is the control tube, V_2 a voltage amplifier, and V_3 the V-R tube. When output voltage rises higher than the value for which the potentiometer is set, the potential difference between grid and cathode of V_2 increases. The increase of V_2-grid voltage causes more current to flow through R_1, an increase in the grid bias on V_1, an increase in V_1 plate resistance, and restoration of d-c

[1] Sulzer, P. G., Stable Electronic Voltage Regulator, *Electronics*, **23**, 162, December, 1950.

output voltage back to normal. Voltage stability of 0.1 per cent or better is achieved in this manner. Ripple is attenuated by this regulator to a degree comparable with output-voltage stability. Amplifier tubes are connected in cascade to obtain still closer regulation.

Control of large amounts of power is accomplished by thyratrons or ignitrons. These devices generally block a-c voltage applied to the anode during a nonconducting period of each cycle, and then suddenly permit current to flow during the remainder of the cycle or half cycle. Current wave form in the load is similar to that of i_c in Fig. 65, where nonconducting periods are 0 to θ_1 and π to θ_2. Hence, ripple in the load increases as θ_1 increases.

An electronic voltage reference and thyratrons are used together in the circuit of Fig. 71 to form an exciter field regulator. X_1 and X_2 are saturable reactors. Regulator tube V_1

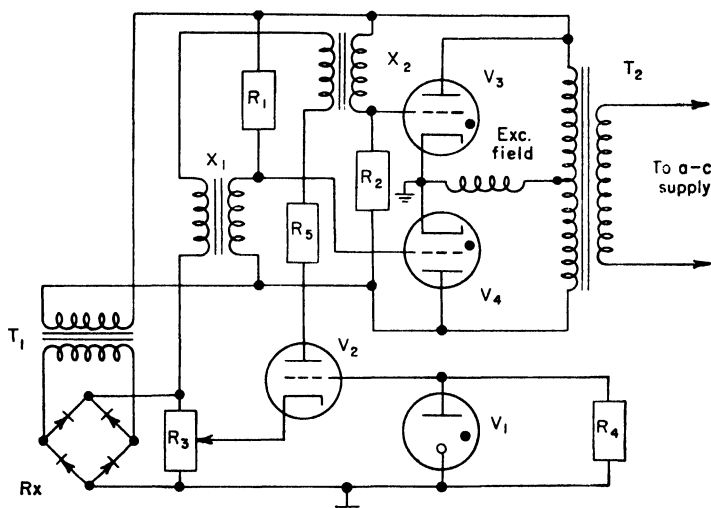



FIG. 71. Thyratron control of generator excitation.

maintains V_2 grid at constant potential. An increase of a-c voltage on T_2 and T_1 causes the cathode potential of V_2 to increase, and reduces its plate current. This reduces the saturation of X_1 and X_2, which in turn shifts the grid-cathode a-c potentials on V_3 and V_4 and thus reduces the generator field excitation. With restoration of normal a-c supply voltage (preset by R_3) the regulating action ceases.

37. Magnetic Regulators.[1] Magnetic amplifiers may be used in place of thyratrons for controlling voltage in the circuit of Fig. 71, but diodes are necessary for d-c output. If these diodes are used in self-saturating circuits the amplifying, and hence the regulating, action is increased. Magnetic references have been developed to replace the V-R tube with gains in reliability and ruggedness. These magnetic devices are frequency-sensitive unless special circuits are used.

A magnetic regulator for controlling a-c line voltage is shown in Fig. 72. This is a resonant-reactor voltage regulator. L_1 is linear. L_2 and C_2 are nearly parallel-resonant at the supply-line frequency and rated voltage. Output current flows through the secondary of L_1 which is of such polarity as to better maintain rated voltage. L_2 is partially saturated at this voltage. If line voltage falls below rated value, less current is drawn by L_2, and the

[1] See Evans, W. G., Saturating Transformer Reference Circuit, *Trans. AIEE*, **74**, 610, 1955. Anderson, F. W., "A Line Voltage Regulator Having Magnetic Amplifier Control," AIEE Paper 56-227.

L_2C_2 combination becomes more capacitive. This capacitive current, drawn through L_1, raises the output voltage. Conversely, if line voltage rises above rated value, the L_2C_2 combination becomes less capacitive and the output voltage falls below the line value. This regulator is practical for ratings up to 10 kva. Output voltage variations of ±1 per

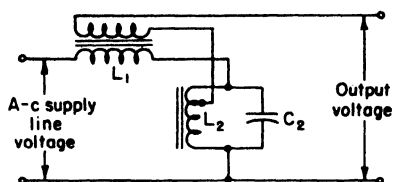



FIG. 72. Resonant-circuit voltage regulator.

cent are obtained with ±10 per cent line-voltage variations in this manner, and with load changes from zero to full load. Constant supply frequency is a condition for resonance at rated voltage; with the good frequency control of modern power systems this condition is generally fulfilled.

CHAPTER 18

RADIO TRANSMITTERS

BY WARREN B. BRUENE[1]

The equipment used for converting electrical power received from a primary source into r-f energy modulated with a signal for transmission by means of electromagnetic waves through space is known as a radio transmitter. Transmitters are usually classified according to the type of signal they transmit, the portion of the spectrum in which they operate, and, frequently, the service for which they are used, for example, medium-frequency broadcast transmitters, high-frequency radiotelegraph transmitters, v-h-f scatter transmitters, v-h-f air-borne communication transmitters, u-h-f television transmitters, etc. (for frequencies, see Chap. 1, Table 1). Transmitters are required to operate in frequency bands or channels assigned by a frequency-allocating authority (FCC) which must consider the service to be performed, the characteristic of the propagation medium, the state of the art, etc. Generally, the low and medium frequencies up through the 540- to 1,600-kc broadcast band depend upon ground-wave propagation which gives reliable continuous area coverage, and the distance depends largely upon the r-f power radiated. Transmitter powers from 250 watts to 500 kw are used in this frequency range. Services which use this frequency range are a-m radiobroadcasting, maritime navigation, and communication in the arctic auroral regions. In the h-f region, 3 to 30 Mc, reflections from the ionosphere cause radio signals to travel long distances with relatively low attenuation. Transmitters ranging from a few watts to 200 kw are used, but powers of 100 watts to 5 kw are most common. Directional antennas providing gains of up to 20 db are employed frequently. The higher-powered transmitters are usually used for multichannel circuits and for short-wave broadcasting to provide adequate reception with simple receivers and poor antennas. The demand for frequency assignments in this part of the spectrum which is useful for long-distance transmission is very great. Those services able to use other frequencies are required to do so. Services such as long-distance communication with aircraft, ships at sea, some point-to-point services which must have independent circuits or which are inaccessible to other means, short-wave broadcast, and the vast military requirements for an independent flexible mobile system are the principal users of the high frequencies. The frequencies in the 30- to 50-Mc region are beyond the frequencies which are reflected by the ionosphere and have been used largely for communication with mobile stations. During the last few years ionospheric scatter propagation has been found useful and is being exploited for reliable continuous-duty circuits over distances of 350 to 1,200 miles. The 30- to 50-Mc region is optimum for this form of propagation. High power on the order of 20 to 50 kw and high-gain antennas are essential. Above 54 Mc are 6-Mc-wide television channels. Power outputs of 5 to 50 kw are usually used with high-gain antennas to achieve effective radiated powers up to 1,000 kw. A great many services use the frequencies above 100 Mc. In general, they use low-power transmitters for general coverage to and from local fixed or mobile stations, and for various navigation services. Microwave stations usually use very low transmitter powers with very-high-gain antennas. An exception is circuits which use tropospheric scatter propagation. These frequently use powers of 1 to 10 kw with high-gain antennas for transmission over distances up to about 350 miles.

[1] Collins Radio Company, Cedar Rapids, Iowa.

The uses and demands for more communications circuits have spurred development of equipment to make use of higher and higher frequencies and also to make better use of the frequency spectrum available.

Many variations of the basic types of modulation are used, and more are continually being developed. Each has its advantages and disadvantages and general field of usefulness. The channel band width required depends upon the type of modulation (a.m., f.m., pulse, etc.), carrier-frequency stability, and the selectivity of receiving equipment. More and more channels are becoming available as the state of the art in each of these fields advances. At present, it appears that single side band (SSB) makes the most efficient use of spectrum space when transmitting a voice signal, and radiotelegraph systems using nearly ideal code modulation and detection[1] are best in this field. Both require very high frequency stability at both transmitter and receiver and hence are restricted to frequencies below 50 Mc until the art of obtaining higher frequency stability improves. Above 50 Mc, frequency modulation is usually used which obtains a signal-to-noise advantage by using more spectrum space. It also has the advantage of transmitter simplicity, which is important in mobile equipment. Another advantage is relative ease of obtaining a wide audio band width for high fidelity. In the microwave region various types of pulse modulation are used to realize signal-to-noise advantages and make the best use of this part of the frequency spectrum.

ELEMENTS OF TRANSMITTERS

Each stage, which includes the circuit about a vacuum tube, is usually described by the function it performs. In addition, various sections of a transmitter are usually referred to according to their function. The more important ones, which are generally similar in many types of transmitters, are exciters, r-f power amplifiers, audio amplifiers, modulators, power supplies, control circuits, and cooling systems.

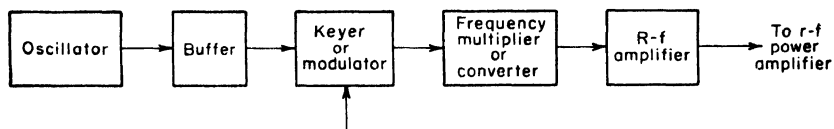



Fig. 1. Block diagram of typical exciter.

1. Exciters. The exciter is usually considered that part of the transmitter which generates the transmitter output frequency and raises it to a power level suitable for exciting the r-f power-amplifier stages. There is a very wide range of exciter requirements and functions; frequently that part of the transmitter which is mounted physically on a single chassis or the transmitter unit which most nearly meets the above definition is called the exciter. The block diagram shown in Fig. 1 illustrates how some of the circuits found in exciters are used.

2. Oscillators. Oscillators provide the source from which the carrier frequency is generated. They must provide r-f voltage precisely at the desired frequency, with a high order of frequency stability and free of hum or noise modulation. The FCC specifies certain limits within which the transmitter frequency must be maintained. For example, m-f transmitters may be allowed a range of ±20 cps from the assigned frequency, while h-f transmitters may be required to be within 0.005 per cent of carrier frequency, which represents ±500 cps at 10 Mc. The limits are established to be about as high as the state of the art permits in a practical economical transmitter design. Limiting the frequency tolerance permits transmitter channels to be assigned closer together and allows use of greater selectivity in the receiver. Some communications systems, such as SSB, require frequency stabilities much greater than this unless automatic frequency control (AFC) is used in the receiver.

There are many factors which affect the frequency stability of an oscillator. Some

[1] Doelz, M. L., Predicted Wave Radio Teleprinter, *Electronics*, December, 1954, p. 166.

of the more important are d-c plate voltage, temperature, humidity, load impedance, and mechanical rigidity. Frequency variations due to varying d-c plate voltage can be minimized by providing a very well regulated and well filtered power supply. Temperature effects are reduced by placing the oscillator, or at least its temperature-sensitive frequency-determining components, in a temperature-controlled oven. Hermetic sealing provides a means of overcoming the frequency variations due to changes in relative humidity of the surrounding air. A buffer amplifier stage will isolate the effects of a varying load impedance from the oscillator. Shielding is also usually desired or required to keep r-f energy from a power-amplifier stage from coupling into the oscillator and affecting its frequency. This shows up mostly when the power amplifier is tuned, keyed, or modulated. The oscillators in most modern transmitters use small receiving-type tubes operating with only a few volts across the oscillating circuit. This demands a high order of circuit isolation, particularly when the oscillator is used in a high-powered transmitter.

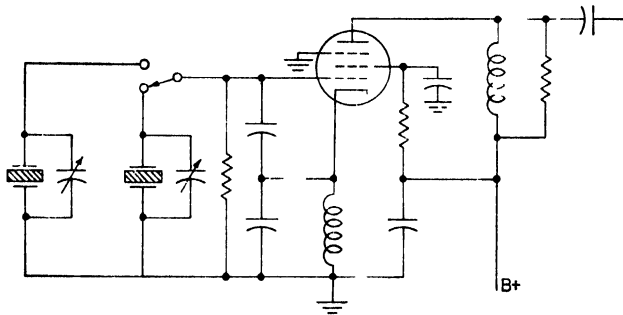



FIG. 2. Popular high-frequency crystal oscillator circuit.

Another effective means of reducing the effects of a small amount of coupling to a power stage is to operate the oscillator on a different frequency than the power stage. In h-f transmitters the oscillator is commonly operated on a subharmonic of the carrier frequency, and frequency multipliers are used to generate the carrier frequency from the oscillator frequency.

Crystal oscillators are normally used because they are capable of much higher frequency stabilities than conventional tuned oscillators. When several frequencies are required, a switch is used to select the appropriate crystal. The quartz-crystal plates are usually mounted in a holder of some type with the frequency marked on it. Oscillation on this exact frequency is not guaranteed in any oscillator circuit other than the standard one in which it was calibrated. The circuit capacitance across the crystal is the important factor. One widely used standard value of oscillator-circuit capacitance is 30 $\mu\mu$f, although other values ranging from half to twice this value have been used. A small trimmer capacitor is frequently placed across the crystal to adjust the total circuit capacitance to the proper value or to trim the crystal into the desired frequency of oscillation. A popular crystal oscillator for transmitters is shown in Fig. 2. It has the advantages that no tuned circuits are required and the output is "electron-coupled" to the oscillating circuit for reduced load-variation effect. The preferred crystal-frequency range for this circuit is 2 to 8 Mc but it may be used well beyond this range.

High-stability operation is more difficult to achieve in tunable oscillators but in some cases their use is necessary. One type of highly developed variable-frequency oscillator (VFO) is shown in Fig. 3. A powdered-iron or ferrite core is moved in and out of the coil to vary the tuning. The coil is wound with a variable pitch so the calibration curve of shaft turns vs. frequency is a straight line and each turn produces an integral frequency range, such as 50 or 100 kc. A final refinement in tuning

linearity is made during the final calibrating procedure during manufacture by means of the cam-type frequency corrector. Each oscillator is individually temperature-compensated to reduce temperature effects and hermetically sealed to minimize effects of relative humidity. One oscillator can tune over a frequency range of a full octave. Best frequency stability can be achieved with this type of oscillator within the frequency range of 450 to 4,000 kc.

3. Frequency Synthesizers. Synthesizer units have been built which can provide fixed frequencies at 1-kc intervals across part or all of the h-f spectrum. A stable crystal oscillator oscillating at 100 kc or at 1,000 kc, for example, is the prime frequency

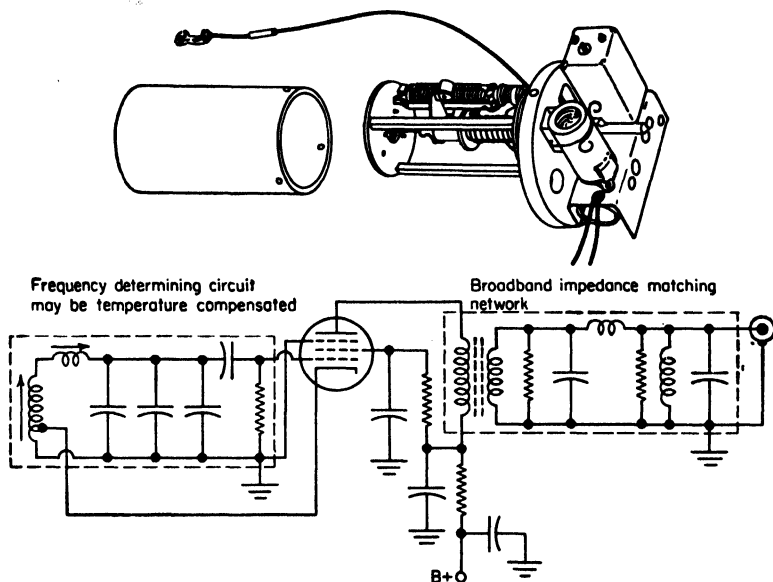



FIG. 3. Collins variable-frequency oscillator.

source from which any of the output frequencies are generated. This is accomplished by using frequency dividers, frequency multipliers, and mixers for addition and subtraction of various selected harmonics and subharmonics of the input frequency. The output frequency has the same frequency stability in per cent as the input frequency because it is derived completely from it. Sometimes the output of a VFO is added to the basic synthesizer circuit to provide interpolation between the fixed output frequencies in such a manner that continuous frequency coverage is achieved with the stability of the VFO contributing only a small amount to the synthesizer output-frequency stability.

Frequency synthesizers, in general, have two serious deficiencies for use as a transmitter carrier-frequency source. One is that it is almost impossible to keep all undesired output frequencies at a satisfactorily low level. The second is that the frequency dividers usually contain an undesirable amount of output-frequency jitter or noise.

4. Stabilized Master Oscillator. A means of overcoming these deficiencies is by use of a stabilized-master-oscillator circuit. The output frequency is derived from a stable master VFO which is further stabilized by trimming its frequency exactly to the desired output frequency by a frequency synthesizer and an a-f-c circuit. Some circuits provide a phase lock of the VFO frequency to the synthesizer output frequency. The stabilized-master-oscillator (SMO) output then has the same frequency stability

as the synthesizer-frequency source, which may be as high as 1 cps at 10 Mc. Figure 4 shows a block diagram of the principal sections of an SMO. The synthesizer may have two or three frequency-setting knobs, although late models show the output frequency directly in kilocycles. The synthesizer controls are coupled to the VFO also, using differential gears so that it is always tuned to be well within the pull-in range of the a-f-c circuit.

FIG. 4. Stabilized-master-oscillator circuit.

5. Frequency Multipliers. Frequency multipliers are used to keep the oscillator frequency off the carrier frequency, to reduce buffer isolation and shielding requirements, and to allow use of an oscillator with a limited frequency range in a wide-range transmitter. For example, a 2- to 4-Mc oscillator can be used with a 4- to 32-Mc transmitter by using frequency multiplications of 2, 4, and 8. This can be accomplished by using three cascaded doubler stages or a quadrupler stage followed by a doubler stage.

R-f frequency-multiplier stages are very similar to class C r-f amplifier stages except that the plate circuit is tuned to select the desired harmonic. The tube is biased well beyond cutoff and a large r-f grid voltage is used to obtain a short angle of plate-current flow which has high harmonic content. About 50 per cent plate efficiency can be realized from a frequency doubler. The efficiency of triplers is somewhat less and the output obtainable from multipliers drops rapidly with higher orders of multiplication.

FIG. 5. Spectrum of frequency-multiplier plate current.

The plate current of frequency multipliers is rich with harmonics, as shown in Fig. 5. The tuned circuits following the multiplier tube must select the desired frequency component and reject others of higher and lower frequencies to a satisfactory degree. This is another reason for using low orders of frequency multiplica-

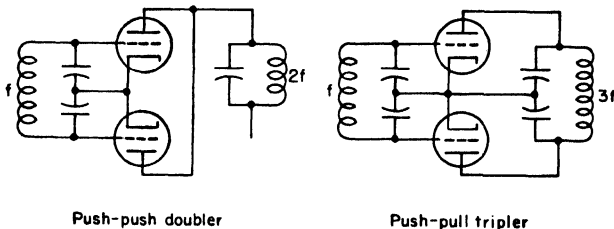



FIG. 6. Multipliers for reduced undesired output.

tion since a tank circuit can better reject frequencies farther removed percentagewise from the desired frequency. It should be recognized that undesired output from the transmitter is possible at any harmonic of the oscillator regardless of the number of multiplier stages or their order of multiplication. Push-push doubler and push-pull

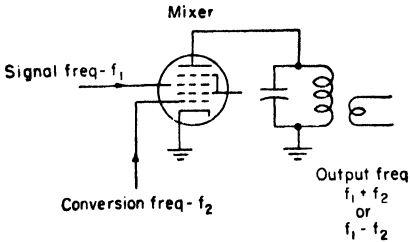



FIG. 7. Typical mixer circuit.

tripler circuits are frequently used to balance out the odd-order and even-order components in their outputs, respectively. Figure 6 shows typical circuits. Cascaded push-pull frequency triplers are commonly used in v-h-f and u-h-f transmitters where high orders of frequency multiplication are required. One example is in f-m broadcast-transmitter exciters.

6. Mixer Exciters. Mixer exciters can raise the frequency of an input signal without changing it in any other way. The output frequency is obtained by addition or subtraction of two frequencies instead of by multiplication, as shown in Fig. 7. Two principal applications of this circuit have been in frequency-shift keyer units and

Ratio of $\frac{A}{B}$ or $\frac{a}{b}$		Order						
		1	2	3	4	5	6	7
0	0	B b	$B \pm A$ $b \pm a$	$B \pm 2A$ $b \pm 2a$	$B \pm 3A$ $b \pm 3a$	$B \pm 4A$ $b \pm 4a$	$B \pm 5A$ $b \pm 5a$	$B \pm 6A$ $b \pm 6a$
1/8	0.125							$7a$
1/7	0.143						$6a$	
1/6	0.167					$5a$		$2B - 5A$ $7A$
1/5	0.200				$4a$		$2B - 4A$ $6A$	
1/4	0.250			$3a$		$2B - 3A$ $5A$		$2b - 5a$
2/7	0.286							$6a - b$
1/3	0.333		$2a$		$2B - 2A$ $4A$		$2b - 4a$ $5a - b$	
2/5	0.400					$4a - b$		$3B - 4A$ $6A - B$
1/2	0.500	a		$2B - A$ $3A$	$3a - b$	$2b - 3a$	$3B - 3A$ $5A - B$	$5a - 2b$
3/5	0.600						$4a - 2b$	
2/3	0.667			$2a - b$		$3B - 2A$ $4A - B$		$3b - 4a$
3/4	0.750					$3a - 2b$		$4B - 3A$ $5A - 2B$
4/5	0.800							$4a - 3b$
1/1	1.000		$2B$ $2A$		$2b - 2a$ $3B - A$ $3A - B$		$3b - 3a$ $4B - 2A$ $4A - 2B$	

Output frequency = $B + A$ or $b - a$.

FIG. 8. Mixer-frequency chart.

in SSB exciters. In each case, the signal is generated in the 100- to 500-kc region and converted to a higher frequency. Two conversion frequencies are usually used to convert the signal to the final carrier frequency in SSB exciters.

A great deal of care must be exercised in the choice of conversion frequencies to avoid undesired frequencies in the exciter output. For example, if the desired output is close to the third harmonic of one of the input frequencies, both will appear in the output and cannot be separated by tuned circuits. These undesired frequencies consist of all the harmonics of the input frequencies plus all the possible sum and difference combinations of all harmonics. The table and chart in Fig. 8 show all mixer products up to seventh order. The upper-case letters indicate that the desired output is the sum of the two input frequencies and the lower-case letters indicate that it is the difference. The amplitude of these various products varies greatly. In general, all second-, third-, and fourth-order products and the stronger fifth- and sixth-order products should be avoided by selecting conversion frequencies so these products can be removed from the desired output frequency with a practical number of tuned circuits. The selectivity of tuned circuits depends upon their circuit Q, source impedance, and coefficient of coupling. In some cases the tuning or tracking error must also be considered. Figure 9 can be used to estimate the attenuation of an undesired signal located a distance from the desired frequency using a single resonant circuit. The value of Q used must be the effective circuit Q, including all loading and source impedance effects and not just the Q of the coil alone.

Balanced-modulator mixers are sometimes used which balance out some undesired products resulting in from 20- up to 50-db improvement if balancing adjustments are provided. Figure 10 shows one type of balanced modulator which balances out the conversion-frequency input and all odd harmonics of it.

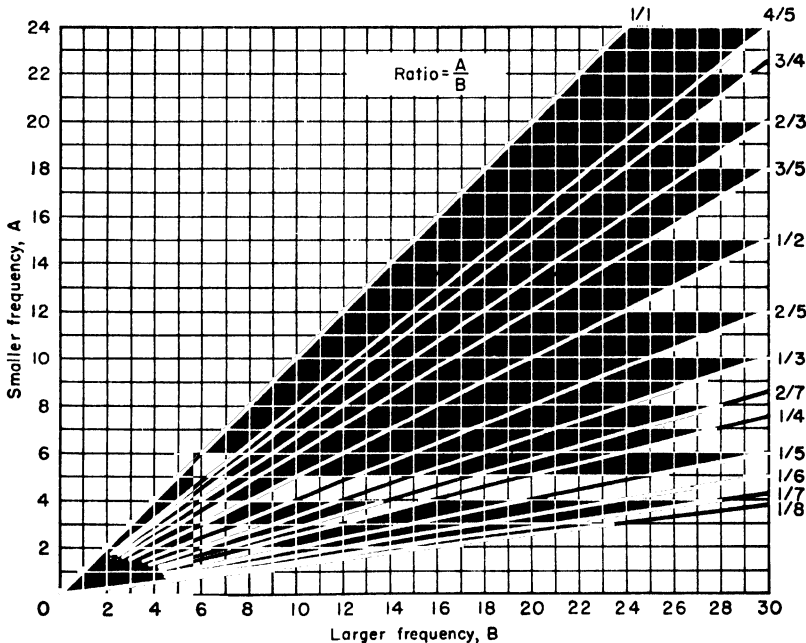



FIG. 8. Mixer-frequency chart. (*Continued.*)

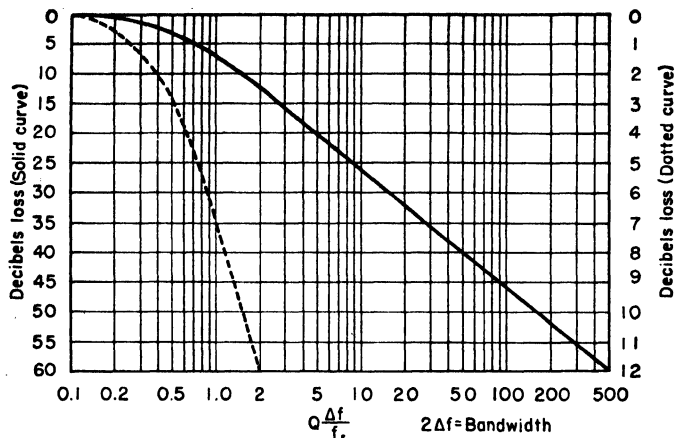



FIG. 9. Generalized selectivity curve, single-tuned circuit.

7. Single-side-band Exciter. In most commercial single-side-band transmitters the SSB signal is generated in an SSB generator at a frequency in the range of 100 to 500 kc. This is raised to carrier frequency by the SSB exciter. Some exciters for h-f transmitters operating in the 4.0- to 30-Mc range convert the SSB generator output to a fixed intermediate frequency of about 3 Mc, as shown in Fig. 11. The signal is heavily filtered at this frequency to remove all undesired mixing products. It is then converted to the desired carrier frequency in one more conversion.

FIG. 10. Balanced modulator.

A frequency scheme suitable for covering the range of 1.7 to 30 Mc is shown in Fig. 12. Two mixing steps are used. It is noted that the two conversion frequencies are obtained from a single oscillator. The first conversion frequency is that of the oscillator and the second is a harmonic of it. This particular frequency scheme has many

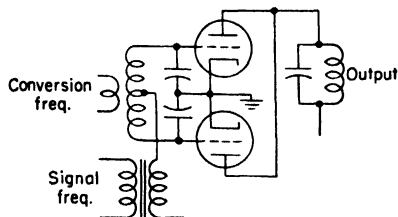
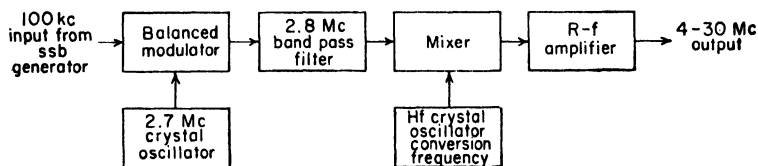



FIG. 11. SSB exciter for 4- to 30-Mc output.

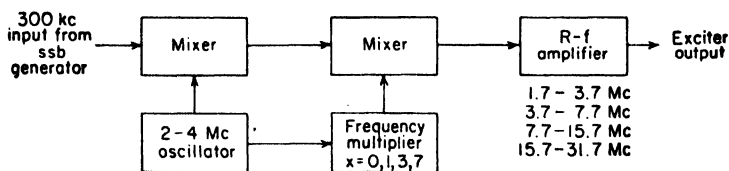



FIG. 12. SSB exciter for 1.7 to 30 Mc.

advantages and was chosen as a good compromise for commercial general-coverage h-f transmitters. Only one stable h-f oscillator is required, and the carrier frequency is easily determined because it is the first, second, fourth, or eighth harmonic of the oscillator less the 300-kc input frequency.

Using two variable conversion frequencies allows the large frequency coverage with a minimum of undesired mixing products near the output frequency.

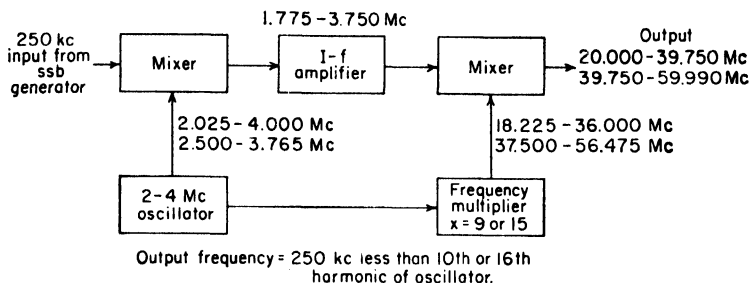



FIG. 13. SSB exciter for 20 to 60 Mc.

An exciter for use in the 20- to 60-Mc range is shown in Fig. 13.

The 250-kc SSB generator output is raised to the desired frequency range in two frequency conversions. Several tuned circuits are required following each conversion to reject the undesired conversion products. The conversion frequencies are synthesized from one very stable standard crystal oscillator to achieve the frequency stability necessary for SSB in this frequency range.

R-F POWER AMPLIFIERS

Tuned circuits are generally used in the input and output circuits of r-f power amplifiers. Electron tubes, which are the power-generating elements, inherently have a significant amount of input and output capacitance. The tubes cannot work efficiently when shunted by low values of capacitive reactance, so that it is necessary to remove this shunting effect. Tuned circuits are used for this purpose. In addition, they are used to provide impedance matching of the tube to its load and also to provide a continuous sine-wave output from amplifiers with tubes which deliver power during only part of the r-f cycle.

FIG. 14. Typical r-f amplifier.

8. Classification. Radio-frequency amplifiers are divided into classes A, B, and C according to the angle of plate-current flow. (Angle refers to the number of degrees of current flow during an r-f cycle of 360 deg.) Class A amplifiers have a continuous plate-current flow and operate over a small portion of the tube's plate-current range as shown in Fig. 15*A*. This class of amplifier is used for amplification of small signals and in linear amplifiers for low distortion. Its efficiency in converting d-c plate-power input to r-f power output is quite low, but this is seldom of major importance where small signals are amplified. Class B amplifiers have their grids biased to near plate-current cutoff, so that plate current flows for approximately 180 deg of the r-f cycle. Amplifiers operated with appreciably more than 180 deg of plate-current flow but less than 360 deg are called class AB amplifiers. Class AB and class B r-f operation is used in linear amplifiers where higher efficiency and maximum output power with low distortion are required. Plate efficiency is in the range of 50 to 75 per cent, depending upon the tube and its operating condition. Typical uses are amplifiers in low-level a-m transmitters and in SSB transmitters. When the grid is biased well

beyond cutoff so that plate current is less than 180 deg the stage is called a class C amplifier. Its principal advantage is high plate efficiency, 65 to 85 per cent. They are usually operated with the r-f plate voltage in the region of saturation so that the r-f output voltage is very closely dependent upon the d-c plate voltage. Varying the d-c supply voltage to the tube at an audio rate by means of a modulator produces an a-m r-f output voltage. Class C amplifiers are used in the power-amplifier stages of radiotelegraph transmitters for c-w and frequency-shift-keyed output, in phase- and frequency-modulated transmitters, and in a-m transmitters which employ high-level plate modulation.

A subscript number is frequently added to the letters indicating the class of operation to indicate whether or not the tube is operated in the positive grid region over

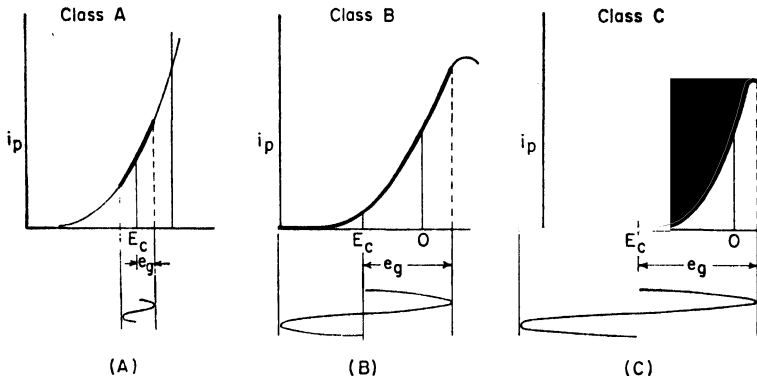



FIG. 15. Plate-current flow in class A, B, and C amplifiers.

part of the cycle. For example, class AB_1 indicates that the grid never goes positive and that no grid current is drawn, while AB_2 indicates that the grid does go positive and that some grid current flows. Because class A amplifiers nearly always operate without grid current, and class C amplifiers nearly always operate with grid current, and because this is assumed unless otherwise designated, the subscripts are seldom used with class A or class C amplifiers.

9. Tube Types. Grid-controlled transmitting-type tubes are used at frequencies up to approximately 1,000 Mc. Klystron-type tubes are popular r-f power amplifiers in the 200- to 3,000-Mc region. Starting in this frequency range and going higher, magnetrons, traveling-wave tubes, etc., are used. Cavity resonators or special resonating circuits are usually required in tubes other than the grid-controlled types. They are usually built in or well defined by the tube manufacturer.

The design of transmitting-type tubes has been undergoing a continual evolution. Thoriated tungsten has almost completely replaced pure tungsten for filament material in high-powered tubes, resulting in a very substantial saving in filament power. The television industry's demand for power tubes in the v-h-f and u-h-f range has resulted in many new developments. In addition to good performance, the demand for reliability and long tube life is increasing. The trend is toward small compact tube design, which is necessary to reach the higher frequencies. External-anode types are increasing since they are more compact, have lower output lead inductance, and allow more efficient cooling. Ceramic insulation for the tube seals, instead of glass, is being used more because it is adaptable to precision tube manufacture, higher-temperature operation, better tube degassing in manufacture, and more rugged construction. Cathode emitters are starting to replace thoriated-tungsten filaments in tubes up to the 1-kw power level. Their high emission capability permits a low-impedance and high-transconductance design which is particularly useful in broad-band circuits. These features, plus the manufacturing precision required in producing tubes with the

small element spacings required for high transconductance and u-h-f operation in general, increase tube cost. The transmitter-design engineer should carefully weigh the factors of reliability, performance, and cost, when selecting a tube for a given application.

Good reliable triode-type tubes are available for nearly every application and power level up into the h-f ranges. More tetrodes and pentodes are being developed to meet the demands for higher power gain and reduction or elimination of the neutralizing requirement. One disadvantage is that screen power must be supplied. Transmitters up to the 50-kw power level now frequently use tetrodes.

10. Calculation of Operating Conditions. The operating conditions of an r-f amplifier tube can be computed rather accurately using Chaffee's method of analysis.[1] The 11-point analysis is suggested because it gives adequate accuracy for almost all engineering purposes and little time is saved by using fewer points. Figure 16 illustrates the principle.[2] A set of constant-current characteristic curves for the tube is required, and sine-wave voltages on the grid and plate elements are assumed. This results in a straight load line on these curves, provided the plate circuit is tuned to resonance (unity power factor). Values are selected for the d-c plate voltage and grid-bias voltage. An operating load line A, Q, A' is chosen, and values of current are read at 15-deg intervals along the load line. These are used in the following equation to compute the d-c component I_{av}.*

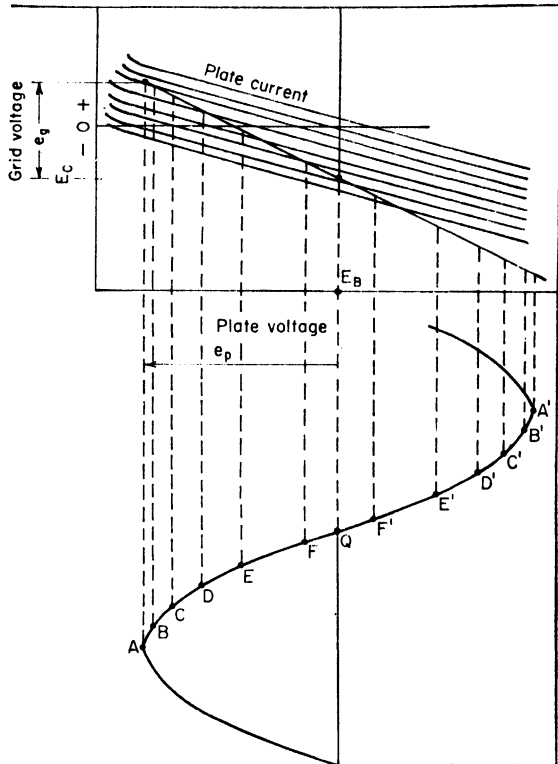



FIG. 16. Graphical-analysis principle.

$$I_{av} = \frac{1}{12}\left[\frac{A}{2} + B + C + D + E + F + Q + F' + E' + D' + C' + B' + \frac{A'}{2}\right] \quad (1)$$

$$I_1 = \frac{1}{12}[(A - A') + 1.93(B - B') + 1.73(C - C') + 1.41(D - D') + (E - E') + 0.52(F - F')] \quad (2)$$

By taking readings of plate, grid, and screen current at the 15-deg points along the load line the d-c and peak a-c components of current can be computed for each element.

[1] Chaffee, E. L., A Simplified Harmonic Analysis, *Rev. Sci. Instruments*, **7**, 384, October, 1936. Sarbacher, R. I., Graphical Determination of PA Performance, *Electronics*, **15**, 52, December, 1942. Eitel-McCullough, Inc., "Tube Performance Computer," Application Bulletin No. 5.

[2] Bruene, Warren B., Linear Power Amplifier Design, *Proc. IRE*, **44**, 1754, December, 1956.

* Subscripts 1 indicate the fundamental component of current or voltage.

These values are then used in the following equations to determine the operating condition rather completely.

$$I_B = I_{p_{av}} \text{ d-c plate current} \tag{3}$$
$$P_{in} = E_B I_B = \text{plate input, watts} \tag{4}$$
$$P_o = \frac{I_{p_1} e_p}{2} = \text{output, watts} \tag{5}$$
$$\text{Eff} = \frac{P_o}{P_{in}} 100 = \% \text{ plate efficiency} \tag{6}$$
$$P_d = P_{in} - P_o = \text{plate dissipation} \tag{7}$$
$$I_{g_{av}} = \text{d-c grid current} \tag{8}$$
$$P_{dr} = \frac{I_{g_1} e_g}{2} = \text{grid driving power, watts} \tag{9}$$
$$P_c = I_{g_{av}} E_c = \text{drive power consumed by bias supply} \tag{10}$$
$$P_{gd} = P_{dr} - P_c = \text{grid dissipation, watts} \tag{11}$$
$$I_{s_{av}} = \text{d-c screen current} \tag{12}$$
$$P_{sd} = I_{s_{av}} E_s = \text{screen dissipation, watts} \tag{13}$$
$$R_L = \frac{e_p}{I_{p_1}} = \text{r-f load resistance} \tag{14}$$

These equations apply to conventional grid-driven circuits. The second and third harmonic components of current may be computed using the following equations.

$$I_2 = \tfrac{1}{12}[(A + A' + C + C' - E - E') + 1.93(B + B' - F - F') - Q] \tag{15}$$
$$I_3 = \tfrac{1}{12}[(A - A') + 1.41(B + D' + F' - B' - D - F) - 2(E - E')] \tag{16}$$

These are used for doubler and tripler computations. Harmonic currents flowing in the plate-tank capacitor, by-pass capacitors, and parasitic suppressors can also be estimated using these equations.

The grid driving power is almost invariably higher than anticipated from the above calculation. Allowance must be made for the grid-tank-circuit losses. In addition, starting in the h-f range, some driving power is fed through the tube due to filament or cathode lead inductance. This includes the internal tube leads, socket, and filament or cathode by-pass capacitors and their leads. This inductance is illustrated in Fig. 17 and the fed-through power can be estimated by the equation

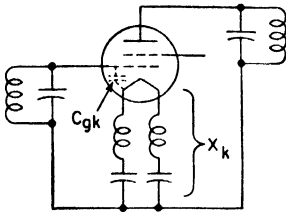



Fig. 17. Filament-lead inductance which causes feed-through power.

$$P_{ft} = \frac{\omega^2 L_k C_{gk} e_g I_{p1}}{2} = \frac{X_k e_g I_{p1}}{2X_{C_{gk}}} \tag{17}$$

The cathode is driven by a voltage equal to the capacitive current flowing from grid to cathode multiplied by the cathode lead reactance. The driving power is this voltage multiplied by the fundamental component of plate current. This increases as the square of frequency and is one reason the older-type tubes with long filament leads became very hard to drive in the upper h-f region. This term can be reduced to zero by choosing a value of filament by-pass capacitor that will series-resonate the total lead inductance.

A net capacitive reactance in the cathode circuit causes plate power to be fed back into the grid circuit, causing a small amount of regeneration.

11. Cathode-driven R-f Amplifier Calculations. Triodes (and occasionally tetrodes) frequently are used in cathode-driven (grounded-grid) circuits as shown in Fig. 18. This circuit has the advantage that neutralization is seldom required because the grid is an effective shield between plate and cathode. The small values of plate-to-cathode capacitance have very little effect on the input signal because the

input impedance is usually only a few hundred ohms. Another advantage is that the minimum circuit input and output capacitance is lower than that of grid-driven triode circuits using broad-band neutralization.

Cathode-driven (or grounded-grid) amplifiers operate a little differently than conventional grid-driven amplifiers. This discussion will cover tetrode tubes, but the considerations for triodes are identical except that operations relating to the screen are omitted.

The calculations are minimized if a set of constant-current curves, which refers all voltages to the No. 1 grid, is available. Frequently they may be obtained from the tube manufacturer if they do not appear on the regular data sheet. The method of using conventional curves for cathode-driven service will be discussed in a later paragraph.

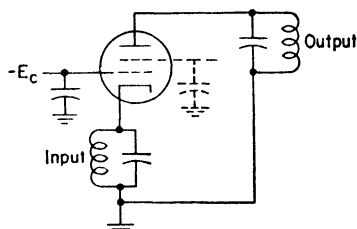



FIG. 18. Cathode-driven power amplifier.

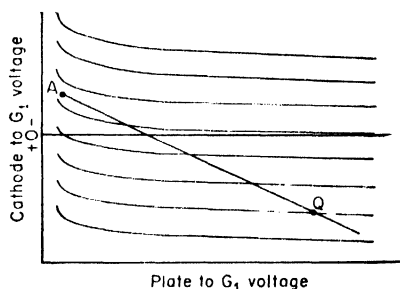



FIG. 19. Load line on constant-current curves for cathode-drive operation.

To calculate the operating conditions, draw a selected load line on the curves for cathode-driven service. Remember that all voltages are referred to the No. 1 grid, but the actual d-c supply voltages normally are measured from the cathode. In practice, the cathode and the negative side of the d-c plate and screen supplies are at d-c ground potential. The point Q on the load line must be at a plate-to-grid No. 1 voltage on the curve, and this is the sum of the d-c plate voltage and d-c grid-bias voltage. For example, 3,000 volts of E_B and 100 volts of E_C gives 3,100 volts of plate-to-grid No. 1 voltage on the curves. The other coordinate of Q is at a cathode-to-grid No. 1 voltage equal to the grid-bias voltage but of opposite sign. The plate-current pulses occur when the cathode is driven in the negative direction. A typical load line on a set of curves is shown in Fig. 19 for class B operation. The d-c and fundamental a-c components are calculated using the method explained in the previous section.

The operating conditions are then calculated as follows:

D-c plate input

$$P_{\text{in}} = E_B I_{p_{\text{av}}} \tag{18}$$

Note that E_B is the d-c plate voltage from plate to cathode and not plate to grid No. 1.

R-f power output

$$P_o = \frac{e_p I_{p_1}}{2} \quad \text{watts} \tag{19}$$

Fed-through power

$$P_{ft} = \frac{e_k I_{p_1}}{2} \quad \text{watts} \tag{20}$$

Plate dissipation

$$P_d = P_{\text{in}} - (P_o - P_{ft}) \tag{21}$$

Driving power

This is composed of three components: the fed-through power, the grid-driving power, and the screen-grid-driving power.

$$P_{dr} = P_{ft} + P_g + P_s = \frac{e_k I_{p_1}}{2} + \frac{e_k I_{g_1}}{2} + \frac{e_k I_{s_1}}{2} \quad (22)$$

Grid dissipation

$$P_{gd} = \frac{I_{g_1} e_k}{2} - I_{g_{av}} E_c \quad (23)$$

Screen dissipation

$$P_{sd} = \frac{I_{s_1} e_k}{2} - I_{g_{av}} E_s \quad (24)$$

Input impedance

$$R_{in} = \frac{e_k}{I_{p_1} + I_{g_1} + I_{s_1}} \quad (25)$$

This varies a little with signal level in a class AB linear amplifier because the grid and screen currents do not vary linearly. At very low signal levels, the grid and screen currents are nearly always zero. In most SSB applications the cathode input impedance is very nearly that due to the fed-through power alone.

Other unknowns of interest may be calculated as discussed in the previous section.

Conventional constant-current curves for grid-driven operation can be used readily for triode cathode-driven amplifiers. A load line is selected as for grid-driven service and all calculations are made. The fed-through power must be added to the power output of the tube to obtain the total power output. Other than this, all the appropriate previous equations apply.

Accurate calculations cannot be made easily for tetrodes using these curves because the screen-to-cathode voltage varies over the cycle and the curves represent only the condition when the screen-to-cathode voltage is fixed. However, reasonably good estimates of power output and fed-through power can be made if all pertinent considerations are kept in mind.

Cathode-driven amplifiers have relatively low gain because of the large amount of fed-through power. When such a stage is plate-modulated, the driver stage must also be modulated so that the fed-through power will be modulated.

TUNED R-F COUPLING CIRCUITS

12. Tank-circuit Considerations. Resonant tank circuits are used in r-f power amplifiers to remove the effects of tube and circuit stray capacitances so that a resistive load can be presented to the tube plate, to maintain sine-wave r-f voltages, to provide a return path for harmonic components of grid and plate current, for selectivity to attenuate harmonic and other undesired frequencies, and for impedance matching. The data sheets for most tetrodes and pentodes list the tube input and output capacitances. Triode-tube data sheets usually merely list the interelectrode capacitances so that the input and output circuit capacitance, including that due to the neutralizing circuit, must be calculated as shown in Secs. 23 to 26. The tube input and output capacitances in cathode-driven operation are equal to the grid-to-cathode and the grid-to-plate capacitances, respectively. To the basic tube and neutralizing-circuit capacitance will be added socket capacitance and the various stray capacitances of leads and associated circuit components. The tuned circuits must be designed to resonate with these fixed values of circuit capacitance across them.

The design of class A voltage amplifiers for transmitters is very similar to that of receiver amplifiers. The tube type and operating condition chosen must be appropriate for the voltage output and gain desired. Tetrodes and pentodes are nearly always

used because they yield a higher gain and need a minimum amount of neutralizing. Their gain can be estimated by using the equation

$$G = g_m R_L \quad \text{or} \quad g_m Q X \tag{26}$$

if g_m at the desired operating d-c grid and plate voltage is known. The value of g_m given on transmitting-tube data sheets is usually measured at some other operating point which is convenient for measurement, so that it is necessary to estimate the g_m in the region of interest from the tube-characteristic curves. Very high voltage gains are theoretically possible when a high-Q tank coil is used in a low-capacitance tank circuit, but circuit stability places a limit on usable gain. The effects of a small amount of grid-to-plate capacitance and stray coupling between the grid and plate circuits increase with the amount of gain used. For high gain, a well-screened tube

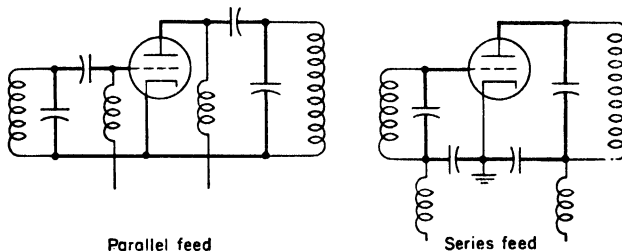



FIG. 20. Essential tank-circuit elements.

with low grid-to-plate capacitance and low-inductance leads along with high g_m should be chosen. Neutralization and good shielding are desirable or necessary in the h-f region and above. The gain of an amplifier can be reduced by loading the output circuit with resistance. Fixed-resistance loading is also used frequently to keep the amplifier gain more uniform when tuning over a band of frequencies by swamping the effect of a varying QX.

It is necessary to maintain a substantial amount of stored energy in the plate tank circuit of class B and class C r-f power amplifiers to maintain approximately a sine-wave r-f voltage. The ratio of stored energy to the energy passing into and out of the circuit during one cycle is expressed as Q. The stored energy in that part of the tuned circuits directly connected to the tube is most important, and a capacitive reactance across the tube elements is desirable for a harmonic-current return path, so that the plate tank circuit Q is often expressed as the ratio of plate load resistance to the shunting-tank capacitor reactance.

$$Q = \frac{R_L}{X_c} \tag{27}$$

Preferred values of Q are between 8 and 20. Lower Q is sometimes used at frequencies below 1 Mc to maintain an adequate band width. Qs as low as 3 in push-pull amplifiers have been used in the 250- to 500-kc region. Care must be used in designing and tuning these circuits because the tube plate efficiency may be impaired. The d-c plate-current dip occurs at the maximum impedance point, which does not necessarily coincide with the unity-power-factor point. This is more noticeable when the element coupled to the load (for example, the tank coil) is tuned instead of a tank capacitor which carries only reactive current.

It is also desirable to maintain an adequate Q in the grid circuit, and a capacitive return from grid to cathode for the grid harmonic currents of class B and C power-amplifier tubes should be provided. Examples are shown in Fig. 20. A means of blocking the d-c grid and plate voltages is required, and the two basic methods (parallel feed and series feed) are shown. The d-c blocking capacitors should have a low

value of reactance, normally a fraction of the tank capacitor reactance. In plate-modulated transmitters the maximum value is limited by the amount of capacitance loading the modulation transformer can stand. A capacitor could be used in series with the tank coil; however, to avoid passing large tank currents through the blocking capacitor, this is seldom done. Coil losses would also be increased because the inductance would have to be increased to overcome the blocking-capacitor reactance. Another type of series feed which overcomes these objections is shown in Fig. 21. D-c, as well as r-f, voltage appears across the tank capacitors, however. Shorter leads from tube to tuning capacitor are possible, giving a better harmonic-current return path and a higher parasitic resonant frequency. Another advantage is that less r-f voltage appears across the series-feed r-f choke, since the impedance at this point is usually much lower than at the plate of the tube.

FIG. 21. Series-fed pi network.

R-f Chokes. The best r-f chokes for use up to 30 Mc, where the plate voltage is only a few hundred volts, seem to be the very common 4-pi chokes rated at 2.5 mh and 100 ma d.c.

Plain solenoid-type r-f chokes are the most satisfactory kind for use across high r-f voltages and over a wide frequency range. They are usually wound with a length of five to ten times their diameter and as small physically as the d-c current and the r-f voltage across the choke will allow. This keeps their distributed capacitance to a minimum. It also keeps the field of the choke as small as possible, so that shields and other objects in the field of the choke do not reduce the series self-resonant frequency and broaden it. R-f chokes must have sufficient inductance for use on the lowest frequency of operation and must have their lowest series self-resonant frequency above the highest frequency of operation.

FIG. 22. Series-fed circuit for increased subharmonic attenuation.

If a suitable compromise cannot be made, it is necessary to use more than one choke to cover the frequency range. They cannot be simply connected in series, however.

Pi-wound commercial chokes often can be used in the amateur bands, but they usually have holes between bands and would burn up at these series self-resonant frequencies.

An interesting choke application in a series-fed circuit useful in doubler stages is shown in Fig. 22. C_1 and L_1 form the plate tank circuit which is tuned to resonance by varying L_1. C_2, for d-c blocking and r-f by-pass, is series-resonant with the equivalent parallel reactance of C_1L_1 at the driving frequency f. $C_2 \cong 1.5C_1$. This places a very low impedance load on the tube at f and results in much higher attenuation of f than would otherwise be obtained. When L_1 is varied for tuning, the two resonant frequencies track across the band.

13. Tank-circuit Design.[1] The basic problem is represented schematically in Fig. 23. In this figure R_L is the required load resistance and Q is the circuit Q. Except for very low values of circuit Q the approximation $X_L = X_C$ is nearly exact. Low Q values are treated later in connection with the L-section design curves. R_L may be

[1] Most of the material in Secs. 13 through 21 is taken from Warren B. Bruene, How to Design R-F Coupling Circuits, *Electronics*, **25**, 134, May, 1952.

determined approximately by the following methods. For class C amplifiers, assume the peak plate-voltage swing to be $0.8E_B$, where E_B is the d-c plate voltage. For high-g_m tubes at maximum plate voltage this figure may be as high as 0.9, and for ordinary triodes with low plate voltages it will be less than 0.8.

Since the power output is known, R_L can be computed from

$$R_L = \frac{e^2 \text{ peak}}{2P} \tag{28}$$

A slightly more accurate value of R_L can be determined from computations of the tube operating conditions using the procedure shown in the previous section.

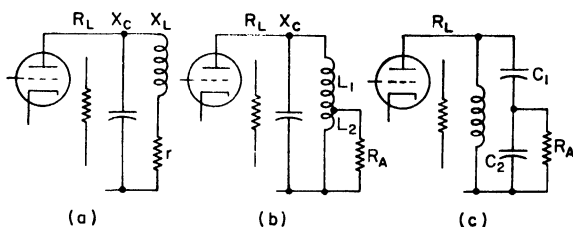



FIG. 23. Basic circuits for matching high-impedance plate circuit to low-impedance load.

With R_L known, X_c can be found by choosing Q and using Eq. (27), $X_c = R_L/Q$. The actual value of capacitance can now be calculated or read from a reactance chart.

The power loss in a simple tank circuit is given by the ratio of circuit Q to coil Q:

$$\% \text{ loss} = 100 \frac{Q_{ckt}}{Q_{coil}}$$

It should be noted that the resistance in the circuit Q includes the r-f resistance component of the inductor.

In direct capacitance-coupled exciter stages, the load resistance is presented directly to the tube. The load on the tube also includes the losses in the tank circuit, which are very appreciable on frequencies where the circuit Q is high, which may be due to high circuit capacitances.

The tap on the tank coil (Fig. 23*b*) may be used to match the tube to any resistive load R_A which is lower than R_L. When the coil is tapped, there is a mutual inductance between the tapped portion and the untapped portion which must be considered. The equation of impedance match for this circuit is

$$\frac{R_A}{R_L} = \left(\frac{L_2 + k\sqrt{L_1L_2}}{L_1 + L_2}\right)^2 \tag{29}$$

where k is the coefficient of coupling. For a given R_A the load on the tube increases as the tap is moved up the coil.

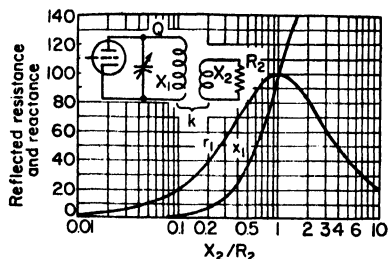
Capacitance division (Fig. 23*c*) may be used to match a low-impedance grid load to a high-impedance driver tube, since a capacitive return to ground is provided for both plate and grid. For a fixed impedance match the ratio of C_1 to C_2 must remain constant. When the circuit Q is high

$$\frac{R_A}{R_L} = \left(\frac{X_{C2}}{X_{C1} + X_{C2}}\right)^2 \tag{30}$$

and the voltage-transformation ratio is

$$\frac{E_A}{E_L} = \frac{X_{C2}}{X_{C1} + X_{C2}} \tag{31}$$

14. Inductive Coupling. When the load is coupled into the plate tank circuit by means of a link or pickup coil, both resistance and reactance will be reflected in series with the tank coil. Figure 24 shows how the loading changes when the number of turns in the pickup coil is varied. The loading is maximum when the reactance of the pickup coil equals the resistance of the load. Under this condition, the ohms of reactance x_1 reflected into the primary equal the reflected resistance r_1.

Fig. 24. Resistance and reactance reflected into tank circuit as per cent of maximum possible reflected resistance.

The reflected reactance drops rapidly as the pickup coil reactance X_2 is decreased. Therefore, if sufficient loading can be obtained with less pickup-coil inductance, the detuning of the plate tank circuit will be less when the coupling is varied. The re-

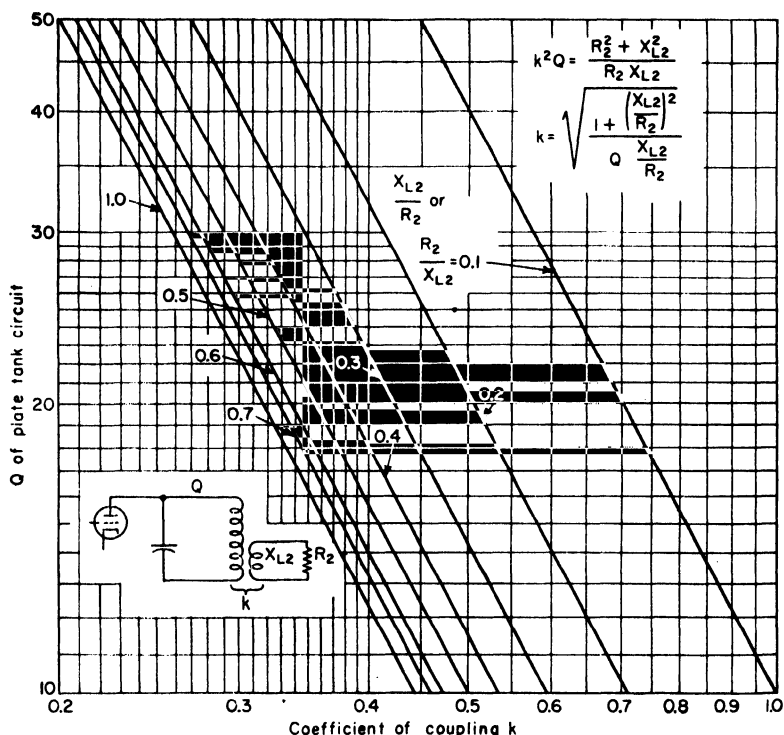



Fig. 25. Relation of k to Q for untuned coupling coil.

flected reactance is tuned out when the final plate current is minimized after any other change is made.

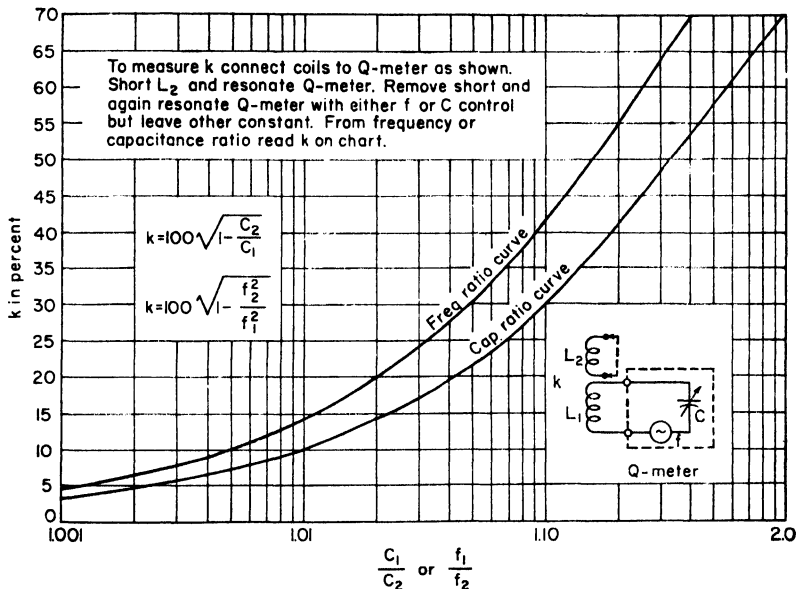
The equations for this circuit are as follows:

$$r_1 = \frac{(\omega M)^2}{R_2^2 + X_2^2} R_2 \quad \text{resistance reflected into primary} \tag{32}$$

$$x_1 = \frac{(\omega M)^2}{R_2^2 + X_2^2} X_2 \quad \text{reactance reflected into primary} \tag{33}$$

$$(\omega M)^2 = k^2 X_{L1} X_{L2} \tag{34}$$

The reflected resistance and reactance are independent of the magnitude of the load resistance and pickup-coil reactance and depend only on the ratio X_2/R_2 and the coupling.

FIG. 26. Use of Q meter to determine coupling coefficient.

For example, if the load R_2 is 600 ohms, the inductance required in the pickup coil may actually be greater than the inductance of the plate tank coil, whereas for a 50-ohm load the inductance of the pickup coil will be very small.

Values of k. Figure 25 shows the relationship between k and Q for various values of X_{L2}/R_2 when the coupling coil is untuned. Since maximum practical values of k are around 0.35, this shows that the pickup coil must be near optimum or the plate-tank-circuit Q will become very high.

In most of these tank-circuit designs it is usually necessary to know the coefficient of coupling obtainable from a given physical arrangement. Sample coils can be wound and the coefficient of coupling quickly determined by use of a Q meter and the coefficient-of-coupling chart shown in Fig. 26. The procedure is given on the chart.

In commercial types of plug-in tank coils, the maximum k where the link is in the center of the tank coil is approximately 0.35 to 0.40. When the link is over the end of the coil it is approximately 0.3. When the link is the same diameter as the tank coil and spaced from the end of it slightly, k is approximately 0.2. When long coils

are coupled together k may be very low since little coupling exists in any but the few turns on each end of the adjacent coils.

There is little reduction of k if the link in the center of the tank coil is wound over it on a larger-diameter form. When the link in the center is wound on a smaller diameter, k decreases proportionally to the reduction in diameter.

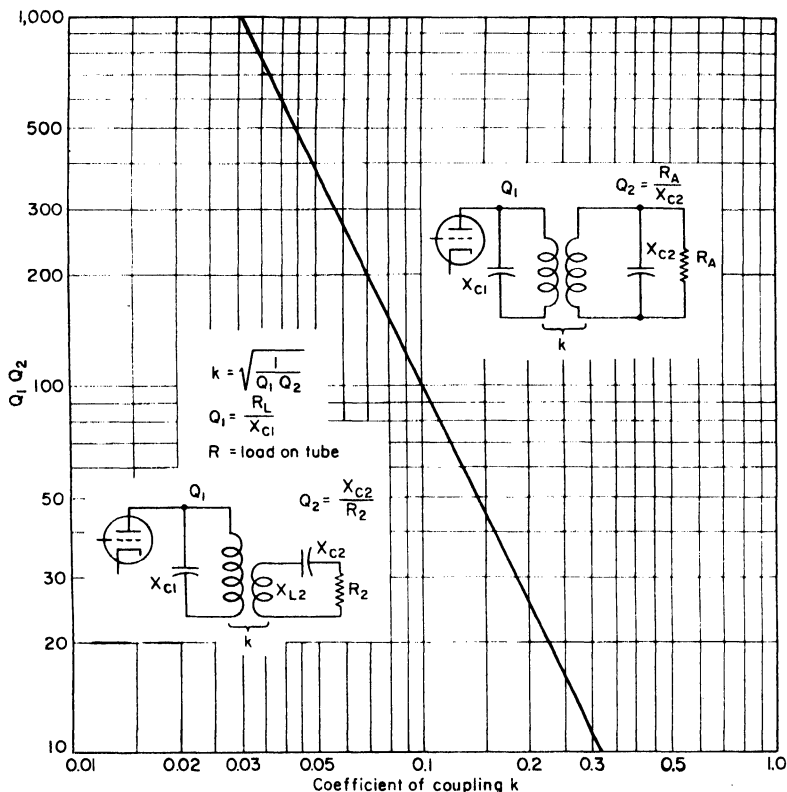



FIG. 27. Relation of k and Q for tuned coupling circuits.

15. Tuned Secondary Circuits. Figure 27 shows the relation of k to tank-circuit Q for circuits having tuned secondaries. The required value is a function of the product of the two circuit Qs. This chart holds true for all values of Q, even those less than unity, and for both series- and parallel-tuned circuits.

For example, if the plate-tank-circuit Q is 10 and the maximum k obtainable is 0.15, what must be the minimum Q of the antenna tank circuit? To solve this, follow the 0.15 line up to the curve (Fig. 27) and then read Q_1Q_2 of 45 at the left. Since Q_1 was given as 10 then Q_2 must be 4.5 or greater.

It is interesting to note that, when one circuit is parallel-resonant and the other series-resonant and both are capacitance-tuned, they can be tuned across a frequency band and still maintain constant Q_1Q_2 and constant k.

16. Link Coupling. Link coupling is often used to couple two circuits together which are physically separated some distance, as shown in Fig. 28. When the transmission line is very short and can be neglected, the effective coefficient of coupling

between the two tank circuits is

$$k_{eff} = \frac{k_1 k_2}{\sqrt{L_3/L_4} + \sqrt{L_4/L_3}} \tag{35}$$

From this equation it is found that the maximum effective coefficient of coupling is realized when the two links have the same inductance. This means that the link inductances should be equal regardless of the inductance of the tank coils or their circuit Qs. When $L_3 = L_4$, the equation reduces to

$$k_{eff} = \frac{k_1 k_2}{2} \tag{36}$$

Fig. 28. Link-coupled tuned circuits. Maximum effective coupling is realized when the two links have the same inductance.

When the length of the transmission line becomes significant and the current or loss in it becomes important, the two links should be designed for matching to the characteristic impedance of the transmission line. The relationships given in Fig. 25 are again used. The adjustments for obtaining a low standing-wave ratio on the line must be made at the load end. A directional wattmeter or some kind of reflected-power indicator is almost necessary for making the proper adjustment. The coupling k_2 and the tuning of L_2 are adjusted for minimum reflected power, resulting in minimum SWR on the line. The tube is loaded by adjusting k_1 and the plate tank circuit, which includes L_1, and is resonated by the usual plate-current dip.

Fig. 29. Harmonic attenuation and output associated with four basic coupling methods.

17. Harmonic Attenuation. Harmonic-attenuation charts for four different types of tank circuits are shown in Fig. 29.

The actual harmonic current generated in an r-f power amplifier is less than the fundamental, of course. In class C amplifiers with 120 deg of plate-current flow, the second harmonic is 3 db less and the third harmonic is about 9 db less. The second harmonic in a class B linear amplifier is 6 db less than the fundamental. The third harmonic would be zero if the tube were perfectly linear. In practice, linear amplifiers are operated more nearly class AB to achieve better linearity, and the second harmonic will then be a little more than 6 db down from the fundamental. These figures added to the attenuation given in the charts give the harmonic level that will appear in the load.

These charts give the maximum attenuation that can be expected, as stray capaci-

tance coupling and other secondary effects may lower these attenuations appreciably. These charts will not be accurate when Q_1 and Q_2 become less than approximately 10 and 3, respectively.

To calculate the harmonic attenuation of any tank circuit proceed as follows:

1. Find the fundamental current through R_L from $I = \sqrt{P/R_L}$.
2. Determine the fundamental current in the load resistance in the same manner.
3. Determine the reactance of all components at the harmonic frequency.
4. Now assume a harmonic current input equal to the fundamental current through R_L and calculate the current that flows in the load resistance. Take short cuts by neglecting resistance terms in each branch as the difference will be only a decibel or two.
5. Use the ratio of fundamental current in the load from step 2 to the harmonic current calculated in step 4 to calculate the decibel harmonic attenuation of the circuit.

18. The L Network. The L network is a very simple circuit; yet it is the most efficient impedance-transforming circuit available. The L-section design chart shown in Fig. 30 is extremely useful in solving nearly all kinds of direct-coupled tank problems.

Example 1: Find the values of X_C and X_L required to match 40 to 200 ohms resistance.

Solution: Find 40 ohms at the bottom of the chart and follow this line up to the intersection of the 200-ohm line extending over from the left-hand scale. The X_C curve (dashed) through this point is 100 ohms and the X_L curve (solid) is 80 ohms, which is the solution.

Example 2: Find the value of X_C that will give a Q of 10 in a tank circuit when the load on the tube is 3,000 ohms.

Solution: The tube load corresponds to R_2 on the chart as it is across the capacitance. Find 3,000 ohms on the R_2 scale and follow across to the diagonal line labeled $Q = 10$. The X_C curve passing through this point is 300 ohms, which is the solution. It can also be noted that the equivalent resistance in series with the tank coil must be 30 ohms.

The L network can be used to match impedances which contain both resistance and reactance. If a low-impedance reactive load is to be matched to a high resistance, the procedure is as follows:

Example 3: Find the values of X_L and X_C required to match $40 + j50$ to 200 ohms resistance.

Solution: Match 40 ohms to 200 ohms as in the first example. The reactive component is in series with X_L, so X_L must be corrected by this amount; so $80 - 50 = 30$ ohms for X_L; X_C remains at 100 ohms. If the reactive component had been $-j50$ then the corrected X_L would be $80 - (-50) = 130$ ohms.

If a low resistance is to be matched to a high-impedance reactive load, the procedure is to convert the reactive load to its equivalent parallel resistance and reactance.

Using lower-case letters for series components and capital letters for equivalent parallel components, equations for this conversion are

$$R = \frac{r^2 + x^2}{r} \tag{37}$$

$$X = \frac{r^2 + x^2}{x} \tag{38}$$

The network reactances are then found on the chart which match the two resistances. The equivalent parallel reactance is combined with X_c to get the corrected value of X_c.

Example 4: Find the values of X_L and X_C required to match 40 ohms resistance to $100 + j100$.

Solution: The equivalent parallel components of $100 + j100$ are $R = 200$ and $X = 200$ ohms inductive. Now find X_L and X_C for matching 40 to 200 ohms. In this case X_C is found to be 100 ohms and this must be corrected to match or tune out the inductive component of the load. Corrected X_C =

$$\frac{100(200)}{100 + 200} = 67 \text{ ohms}$$

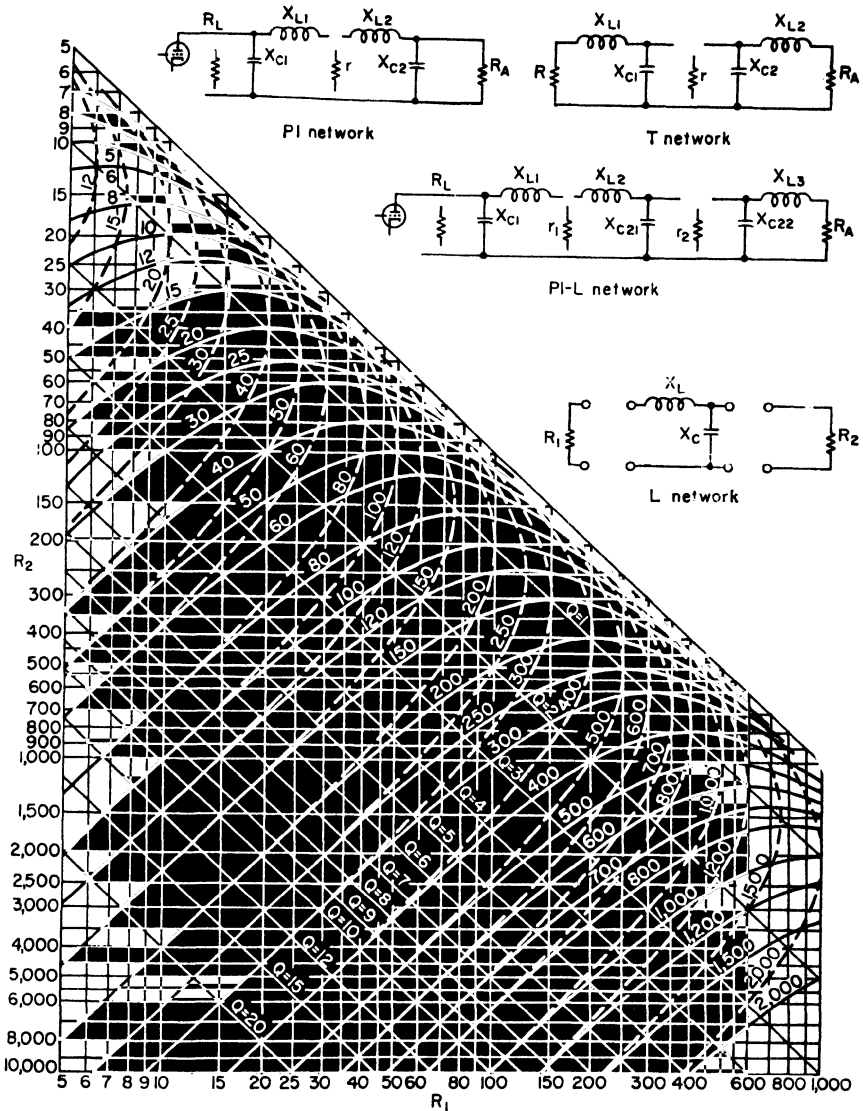



FIG. 30. Design chart for determining optimum component values for L, T, pi, and pi-L networks. Inductive reactances are shown solid; capacitive reactances are dashed.

If the reactive component had been capacitive then it replaces part of X_C and the corrected X_C would be

$$\frac{200(100)}{200 - 100} = 200 \text{ ohms}$$

19. Pi Networks. The pi network can be used to perform the combined functions of a plate tank circuit and an antenna-matching circuit and accomplishes this with a

minimum number of circuit elements. It can match a tube to a wide range of load impedances and still maintain the desired values of plate-tank-circuit Q.[1] They should be used only in unbalanced-grid or plate tank circuits. When used only for impedance matching or harmonic attenuation, they may be used in balanced form.

The conventional pi network has certain impedance-matching limitations and will not always efficiently match a tube to "just any old piece of wire for an antenna." The lowest load resistance that can be matched is approximately[2]

$$R_A = \frac{R_L}{Q^2} = \frac{X_C^2}{R_L} \tag{39}$$

Other limitations depend on the choice of inductance and loading capacitor.

A simple method of determining the values of L and C is to treat the pi network as two separate L sections, each matching into a common imaginary resistance r as shown in Fig. 30. The values of R_L and R_A must be known or found first; then Q is chosen. Now all four capacitive and inductive reactances are found by Fig. 30. Add the inductive reactances together since actually a single inductor is used. Knowing the reactances of these elements, their actual inductance and capacitance can be determined in the usual manner.

Example 5: Find L and C required to match a tube load of 2,000 ohms to a 300-ohm flat transmission line at 7 Mc for a plate-circuit Q of 10.

Solution: Find the intersection of 2,000 ohms and $Q = 10$; then read $X_{C1} = 200$, $X_{L1} = 200$, $r = 20$. Now match this r of 20 ohms to the 300-ohm load. From the intersection of the 20-ohm and 300-ohm lines read $X_{C2} = 80$ and $X_{L2} = 75$. Find total X_L by adding X_{L1} to X_{L2} and get 275 ohms. Therefore $C_1 = 113$ μμf, $L = 6.25$ μh, $C_2 = 285$ μμf.

When the load impedance contains a reactive component, it must be converted to its equivalent parallel resistance and reactance. Then find the values for matching to this equivalent parallel resistance. Correct X_{C2} to tune out the equivalent parallel load reactance and the pi-network values are solved.

20. T Networks. Simple T networks are often used to match one low impedance to another because the values of inductance and capacitance required are of more practical sizes. Also they are often inserted in series with a low-impedance transmission line to provide additional harmonic attenuation. In broadcast-station service they are designed by choosing a suitable value of fixed capacitance and using adjustable taps on the two coils for proper matching. The T-network problems are also easily solved by considering them as two L sections (Fig. 30) and then combining the capacitances.

When using T networks for impedance matching only, keep the Q of the sections as low as practical to avoid unnecessary circuit losses. For harmonic attenuation, use Qs of around 5 or 6 in each L section. Higher Qs do not give much more harmonic attenuation but do increase the losses proportionally.

21. The Pi-L Networks. This circuit is becoming increasingly popular for the final tank circuit when the load is a coaxial transmission line. It makes economical use of the components required and in general provides more harmonic attenuation than other circuits and less circuit loss. The second harmonic output from a class C amplifier stage will be approximately 50 db down from the carrier.

To determine the component values of this circuit, break it down into L sections shown in Fig. 30 and proceed as above.

NEUTRALIZATION

The undesired effects of grid-to-plate capacitance C_{gp} of an r-f amplifier tube can be greatly reduced by using a circuit which balances out or neutralizes these effects.

[1] Bruene, W. B., Pi-network Calculator, *Electronics*, **18**, 140, May, 1945.
[2] Note that R_A is the actual load and R_L is the load needed by the tube for efficient power transfer.

22. Capacitance Effects. The effect of C_{gp} on the effective tube input capacitance is given by

$$\text{Input capacitance} = C_{in} + C_{gp}(1 + A\cos\theta) \tag{40}$$

where C_{in} = tube input capacitance
C_{gp} = grid-to-plate capacitance
A = voltage amplification from grid to plate
θ = phase angle of plate load

In an unneutralized tetrode amplifier with a C_{gp} of 0.25 $\mu\mu$f and a voltage gain of 30, a calculation shows that 7.5 $\mu\mu$f is added to the input capacitance. This may not cause serious detuning at medium r-f frequencies but at high frequencies it may, depending upon the amount of total tuning capacitance and the circuit Q. Triode stages must be neutralized because, with 20 $\mu\mu$f of C_{gp} and a gain of 7, the increase in input capacitance is 160 $\mu\mu$f, which cannot be tolerated. This is particularly true if the tube voltage gain varies with modulation as in a grid-, screen-, or plate-modulated amplifier. The reactance variation with modulation would cause very severe r-f phase shift over the modulation cycle.

The input resistance of the grid is also affected by this grid-to-plate capacitance. It is given by the equation

$$\text{Input resistance} = \frac{1}{C_{gp}(A\sin\theta)} \tag{41}$$

This resistance is in shunt with the grid current loading, grid-tank-circuit losses, and driving-source impedance. When the plate circuit is inductive, there is energy transferred from the plate to the grid circuit through C_{gp} (positive feedback), which introduces negative resistance in the grid circuit. When this shunt negative resistance across the grid circuit is lower than the equivalent positive resistance of the grid loading, circuit losses, and driving-source impedance, the amplifier will oscillate.

When the plate circuit is in resonance ($\theta = 0$), the input resistance due to C_{pg} becomes infinite. As the plate circuit is tuned to the capacitive side of resonance, the input resistance becomes positive, and power is actually transferred from the grid to the plate circuit. This is why the grid current in an unneutralized tetrode r-f amplifier varies from a low value with the plate circuit on the low side of resonance to a high value on the high side of resonance. Grid current, of course, is a common indication of r-f voltage on the grid, but it is the voltage change which is of primary concern. In a pentode or tetrode amplifier operating class A or AB_1, the effect of plate-to-grid feedback can be observed by placing an r-f voltmeter across the grid circuit and observing the change as the plate circuit is tuned through resonance.

If the amplifier is overneutralized, the effects reverse so that with the plate on the low side of resonance the grid voltage is high, and on the high side, it is low. One useful "rule-of-thumb" method of checking the neutralization of a stage (assuming it is nearly correct to start with) is to tune both grid and plate circuits to resonance. Then, observing the r-f grid voltmeter, tune the plate circuit to the high-frequency (low-C) side of resonance. If the indication goes up, more neutralization is required, and if it goes down, less is required. This indication is very sensitive in a neutralized triode class C amplifier, and correct neutralization exists when the grid voltage (or current) peaks at the plate-current dip. In tetrode power amplifiers this indication is less pronounced. Sometimes in a supposedly neutralized tetrode amplifier, there is practically no change in grid voltage as the plate circuit is tuned through resonance, and in some amplifiers it is unchanged on one side of resonance and drops slightly on the other. Another observation sometimes made is a small dip in the center of a broad peak of r-f voltage. These various effects are probably caused by coupling from the plate to grid circuit through other paths which are not balanced out by the neutralizing circuit used.

The grid-reaction method of neutralization is preferred because it can be used with any type of neutralizing circuit. Some neutralizing circuits contain a small amount

of feedback and adjusting for minimum feed-through power does not give the correct neutralizing adjustment. If minimum feed-through power is essential, then this must be considered the selection of the neutralizing circuit.

23. Coil Neutralization of Triodes. Coil neutralization has been used frequently in high-powered amplifiers in the broadcast-frequency and in the lower h-f ranges. A coil L_n is placed across C_{gp} and resonated with it to balance out its effect, as shown in Fig. 31. A d-c blocking capacitor must be connected in series with the neutralizing coil. Sometimes a small variable capacitance placed across the coil is used to make the neutralizing adjustment if a means of trimming the inductance to the correct value is not convenient. This circuit is capable of excellent neutralization, although perfect neutralization would require an infinite-Q neutralizing coil. There is no r-f feedback and no fed-through power. It can be adjusted by removing d-c plate voltage and adjusting for minimum r-f voltage across the plate circuit, as well as by the grid-reaction method. Another feature is that the grid input capacitance is equal to just

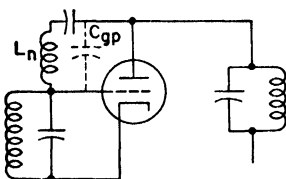



Fig. 31. Coil-neutralizing scheme.

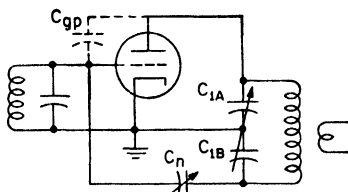



Fig. 32. Plate neutralization.

the grid-to-filament capacitance plus strays. The tube output capacitance is essentially all stray capacitance. The tube is neutralized at only the frequency where $X_L = X_{C_{gp}}$. For this reason this circuit is used principally in transmitters that operate on one fixed frequency.

24. Plate Neutralization. Figure 32 shows a neutralizing circuit capable of good neutralization with very little feed-through. A split-stator tuning capacitor is used and C_n is approximately equal to C_{gp}. The voltage at the lower end of the plate tank circuit is approximately equal and opposite in voltage to that on the plate side. Proper adjustment of C_n then balances the effect of C_{gp}. The equation for neutralizing balance is

$$\frac{C_n}{C_{gp}} = \frac{C_{1B}}{C_{1A}} \tag{42}$$

It is essential that the stray capacitances across C_{1A} and C_{1B} be equal to maintain neutralization accurately when the tank capacitor is tuned to another frequency requiring a different capacitance setting. If necessary, a small padding capacitor can be added to balance them.

Theoretically, the voltage at the lower side of the tank circuit is not exactly 180 deg out of phase with the plate voltage unless the tank-circuit Q is infinite. For this reason low values of plate-tank-circuit Q must be avoided when using plate neutralization, or the power output must be taken directly from the tube plate with an additional output network.

This circuit is of the broad-band type in that the neutralizing adjustment remains constant when tuned to any frequency within a broad range of frequencies. It should be understood that the tube is neutralized at just the one frequency to which the amplifier is tuned. Neutralization is lost when the plate circuit, which forms part of the neutralizing circuit in this case, is tuned off resonance. The stage remains stable and does not oscillate because the gain of the amplifier also becomes very low when the plate tank circuit is off resonance. Particular note should be made of the

fact that C_{gp} is not neutralized at parasitic frequencies. In fact, in some low-frequency parasitic cases, the neutralizing capacitance is effectively in parallel with C_{gp}.

Plate neutralizing has the disadvantage that the power loss in the plate tank coil is about twice the losses in a single-ended plate circuit because twice the inductance is required. The large circulating current and energy storage in the neutralizing half of the plate tank circuit have little value other than supplying a neutralizing voltage.

25. Grid Neutralization. A grid-neutralized circuit is shown in Fig. 33. A balanced grid tank circuit is used instead of a balanced plate tank circuit. Negative feedback is inherent in this circuit. In some applications, such as class B linear amplifiers, this may be acceptable or desirable, while it may be very undesirable in

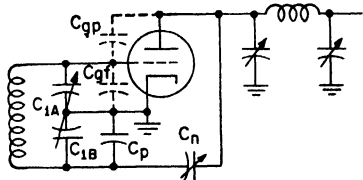



Fig. 33. Grid neutralization.

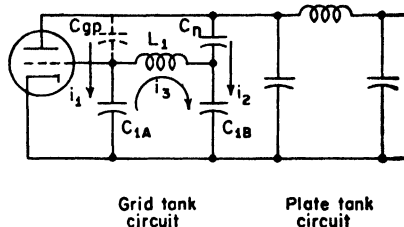



Fig. 34. Feedback inherent with grid neutralization.

some amplitude-modulated amplifiers. Figure 34 shows this feedback more clearly. For neutralization

$$\frac{C_n}{C_{gp}} = \frac{C_{1B}}{C_{1A}}$$

where C_{1A} and C_{1B} include all tube and stray capacitance across them. The feedback voltage is the plate voltage e_p stepped down by the capacitance-voltage division of C_{gp} and C_{1A}.

Another point of interest in this circuit is that the voltage across C_{1B} is higher than across C_{1A}. This is because the current i_2 in the neutralizing branch adds to the circulating grid tank current i_3 while the plate-to-grid capacitive current subtracts from i_3 in C_{1A}. The voltage drop across each capacitor is equal to its reactance times the current flowing through it.

This grid-neutralizing circuit should have its final adjustment made under normal full-power operating conditions using the grid-reaction method. The grid current (or voltage) should peak when the plate is at resonance and fall off symmetrically when tuning it to each side of resonance.

The effective input capacitance across the grid tank circuit due to the tube interelectrode capacitance and the neutralizing circuit is

$$C_{\text{in}} = \tfrac{1}{2}C_{gf} + \tfrac{1}{2}C_{gp}$$

The equivalent value across each half of the grid tank is twice this value. The output capacitance across the plate tank circuit is

$$C_{\text{out}} = C_{pf} + 2C_{gp}$$

Grid neutralization has the advantage that any type of single-ended output network can be used. This minimizes power loss and uses the fewest components.

26. Push-pull Cross Neutralization. Cross neutralization is nearly always used in push-pull triode amplifier circuits. It is ideal in that all voltages and currents remain in balance and there is no feedback. The two neutralizing capacitors C_{n1}

and C_{n2} (in Fig. 35) should be set to the same value and adjusted together. The adjustment can be made with the plate voltage off by exciting the grids and adjusting for minimum voltage in the plate circuit. This can be sensed with a neon bulb, a flashlight bulb in series with a pickup loop, or any sensitive r-f detector. The grid-reaction method can also be used.

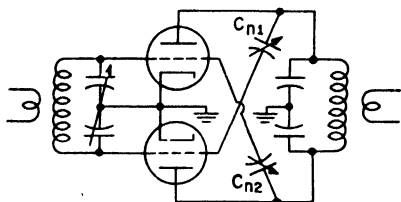



FIG. 35. Push-pull cross-neutralized circuit.

The neutralizing capacitors add some capacitance across the input and output circuit. The equivalent grid-to-grid input capacitance is

$$C_{\text{in}} = C_{pg} + \tfrac{1}{2}C_{gf}$$

From plate to plate the capacitance is

$$C_{\text{out}} = C_{pg} + \tfrac{1}{2}C_{pf}$$

The superior performance of cross neutralizing was undoubtedly one of the principal reasons for its high popularity for many years.

27. Triode Neutralizing-circuit Design Considerations. Usually a good bit of engineering is required to obtain best performance of any of these three wide-band neutralizing circuits for triodes. Low-inductance leads and maintenance of balance in balanced circuits over the frequency range are essential. A point often overlooked is that these neutralizing circuits can contain resonances which may cause parasitic oscillations. Many of these considerations are related to the operating frequency, tube size and construction, size of components, value of C_{gp}, and the circuit chosen. Class C amplifiers usually do not need to be neutralized as accurately as class B linear amplifiers. In general, good neutralization is easy to obtain up to 5 or 10 Mc but care must be used above 10 Mc in medium- and high-powered stages. Conventional neutralized triode stages are seldom used above 20 or 30 Mc. Cathode-driven (grounded-grid) circuits are usually used in this region because the neutralizing requirement is practically eliminated by the screening action of the control grid and the low input impedance.

There are many other neutralizing circuits but they are falling into disuse because, in general, they are not easily adapted to rapid frequency changing.

28. Neutralization of Tetrode Tubes. In addition to high power gain, tetrodes have very low values of C_{gp}. Usually it is low enough to prevent oscillation when the stage is properly tuned and loaded. Shielding is necessary to isolate the grid and plate circuits so that C_{gp} is the largest remaining coupling element. It is generally desirable to neutralize tetrodes, however, to reduce the reaction of plate tuning on the grid drive. This is particularly true in high-gain linear amplifiers operating class AB.

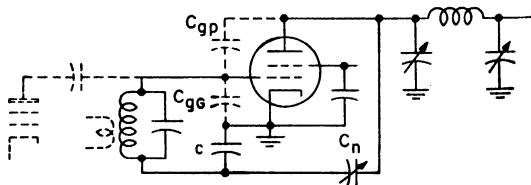



FIG. 36. Bruene neutralizing circuit for tetrodes.

Any of the broad-band neutralizing circuits described for use with triode tubes can also be used with tetrodes. The neutralizing capacitor C_n is very small and a small post mounted near the tube provides sufficient capacitance to the tube plate. The balanced tank circuit used for grid or plate neutralization of triodes is avoided in the circuit shown in Fig. 36.[1] The split-stator tuning capacitor and the extra coil loss are avoided. The out-of-phase voltage is obtained across C, which also serves as the grid-tank by-pass capacitor. It is similar to conventional grid neutralizing except that the out-of-phase voltage across C is much less than the grid voltage across C_{gG}. In

[1] Bruene, Warren B., Single-ended Tetrode Final, *CQ*, August, 1950, p. 11.

the illustration C_{gG} represents all capacitance from grid to ground which includes the tube-input capacitance, all stray lead and component capacitance to ground, and the output capacitance of the driving stage if it is capacitance-coupled. The relationships for neutralization are

$$\frac{C_n}{C_{gp}} = \frac{C}{C_{gG}} \tag{43}$$

Typical values of C lie in the region of 200 to 2,000 $\mu\mu$f and C_n is then usually about ten times C_{gp}.

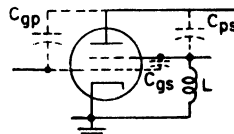
A small amount of negative r-f feedback is inherent in this circuit; so adjusting for minimum r-f feed-through does not give proper neutralization, although it can be used to get into the region of proper adjustment. The grid-reaction method gives good results, particularly in class C amplifiers. Some tetrode amplifiers have very little plate-current dip at resonance when properly loaded. This is particularly true of class AB_1 tetrode amplifiers. Also, the grid-voltage peak may be difficult to distinguish. A more sensitive indication is the peak in power output or drive to the next stage at the plate-current dip. Adjusting the neutralizing capacitance to make the power-output peak coincide with the plate-current dip usually results in a very accurate neutralizing adjustment. The peak in d-c screen current is usually a more sensitive indication of plate tuning than the plate-current dip. This can be used if a meter is provided in the screen circuit.

FIG. 37. Tube elements involved in plate-to-grid coupling.

The screen by-passing lead impedance has a pronounced effect on the neutralization of many power tetrodes.[1] In fact, they may become self-neutralized at some frequency in the v-h-f region. Above this frequency they are overneutralized and require the addition of plate-to-grid capacitance. This is characteristic of power tetrodes using rods to support the screen. Figure 37 illustrates how the tetrode interelectrode capacitances (shown dashed) are involved with inductance L, which is in the screen by-pass connection. R-f voltage on the plate of the tube causes a current to flow through C_{ps} and L. A voltage appears on the screen that is out of phase with the plate voltage. The grid is coupled to this r-f voltage by C_{gs}. At some frequency this will just neutralize the effect of C_{gp}. The effective value of L can be reduced by partially series-resonating it with a variable capacitor to obtain neutralization. In general, it is better to keep L at a minimum by using a very-low-inductance screen by-pass capacitor and short connecting leads. Above the self-resonant frequency a small amount of capacitance shunted across C_{gp} will produce neutralization. The neutralization adjustment changes with operating frequency since an inductive element is involved.

Many of the more recently developed transmitting-type tubes use a cone to connect the screen to a ring terminal. This greatly reduces the internal screen-lead inductance. Screen by-pass capacitors of very low inductance can be constructed which form a ring around the tube. Silvered-mica rings generally form the dielectric for the smaller transmitting-type tubes. This reduces the series screen inductance to such low values it is insignificant.

By-passing the ring-seal screen terminals of the larger power tetrodes is more difficult, however. Undesirable resonances may appear in the by-pass capacitor because of its large physical dimensions. The wavelength of a standing wave in dielectric material is shorter than the wavelength in air by the square root of the dielectric constant. For example, one wavelength at 200 Mc in air is about 5 ft. In a dielectric with a dielectric constant of 25, the wavelength is only 1 ft. Resonance in the frequency range where parasitic plate and grid resonances occur must be avoided.

One method of circumventing this problem is to connect the ring screen terminal

[1] Eitel McCullough, Inc., "The Care and Feeding of Power Tetrodes," Application Bulletin No. 8, 1951.

directly to the chassis deck all around its periphery. This provides the best possible isolation of the tube input and output circuits, as shown in Fig. 38. The neutralizing holds over a wider frequency range and parasitic tendencies are greatly reduced. The filament is operated at a negative voltage equal to the value of the d-c screen voltage. The grid-blocking and the filament-by-pass capacitors have this additional value of voltage across them; otherwise their requirements remain the same.

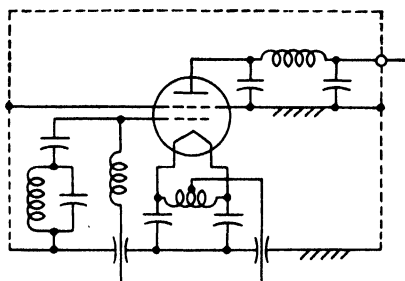



Fig. 38. Grounded-screen power amplifier.

MODULATION

29. Types of Modulation. An r-f wave must vary in some manner to carry intelligence. The process of varying the wave in accordance with the signal to be transmitted is called modulation. Some types of r-f signals do not use a carrier, since it actually contains no information. Single side band is the conversion of an audio signal to an r-f frequency either with or without inversion. This is identical to just one side band of an a-m signal. Pulse modulation is limited to microwave frequencies, since it uses much more spectrum space. It is very useful in microwave-relay systems because it has a good signal-to-noise ratio and the pulse can be reshaped at each relay station, which substantially removes all noise.

Fig. 39. Effect of key-click filter on keyed wave shape.

Certain letters with subscripts have been adopted to designate the various types of transmitter modulation or r-f emission. Table 1 is reproduced from the FCC, General Rules and Regulations, Part 2—Frequency Allocations and Treaty Matters.

30. Keying. Conventional on-off keying has generally been referred to as "c.w." although the official designation is A_1. It is used when manual sending and receiving methods of code transcription are employed.

Undesirable keying characteristics are chirps, clicks, and backwave. Key clicks are caused by turning the carrier on and off too fast as the key is closed and opened. A continuous pair of side-band frequencies extends out from the carrier up to 100 kc in extreme cases. A key-click filter introduces a time constant to slow down the rate of carrier rise and fall as shown in Fig. 39, which greatly reduces the clicks and limits them to a narrow region near the carrier frequency. The method of shaping the keyed wave depends upon the keying circuit used. Frequency multipliers and class C amplifiers tend to square up the wave shape, since they operate in the region of saturation. This effect varies with changes in gain which may be due to operating on various frequencies. This reduces the effectiveness of wave shaping in an early keyed stage and also lengthens the mark period as shown in Fig. 40.

Fig. 40. Effect of class C amplifier on keyed wave shape.

Keying and wave shaping are sometimes done in the final power-amplifier stage to

Table 1. Modulation Symbols

Type of modulation of emission	Type of transmission	Supplementary characteristics	Symbol
1. Amplitude	Absence of any modulation	....	A0
	Telegraphy without the use of modulating audio frequency (on-off keying)	....	A1
	Telegraphy by the keying of a modulating audio frequency or frequencies or by the keying of the modulated emission (special case: an unkeyed modulated emission)	....	A2
	Telephony	Double side band, full carrier	A3
		Single side band, reduced carrier	A3a
		Two independent side bands, reduced carrier	A3b
	Facsimile	....	A4
	Television	....	A5
	Composite transmissions and cases not covered by the above	....	A9
	Composite transmissions	Reduced carrier	A9c
2. Frequency (or phase) modulated	Absence of any modulation	....	F0
	Telegraphy without the use of modulating audio frequency (frequency-shift keying)	....	F1
	Telegraphy by the keying of a modulating audio frequency or frequencies or by the keying of the modulated emission (special case: an unkeyed emission modulated by audio frequency)	....	F2
	Telephony	....	F3
	Facsimile	....	F4
	Television	....	F5
	Composite transmissions and cases not covered by the above	....	F9
3. Pulsed emissions	Absence of any modulation intended to carry information	....	P0
	Telegraphy without the use of modulating audio frequency	....	P1
	Telegraphy by the keying of a modulating audio frequency or frequencies, or by the keying of the modulated pulse (special case: an unkeyed modulated pulse)	Audio frequency or frequencies modulating the pulse in amplitude	P2d
		Audio frequency or frequencies modulating the width of the pulse	P2e
		Audio frequency or frequencies modulating the phase (or position) of the pulse	P2f
	Telephony	Amplitude-modulated pulse	P3d
		Width-modulated pulse	P3e
		Phase (or position) modulated pulse	P3f
	Composite transmissions and cases not covered by the above	....	P9

avoid this effect. Chirps are caused by a small but rapid change in frequency as the carrier is keyed on or off. This is particularly troublesome when the oscillator itself is keyed. Also the keying of a later stage may change the oscillator supply voltages or the load impedance on the oscillator, resulting in a chirp. Wave shaping lengthens the time of frequency change and makes the chirp more noticeable. With no wave shaping the frequency change is so rapid it sounds like a click. Isolating the keyed stage with a buffer stage and using well-regulated oscillator supply voltages can reduce chirp to a negligible value.

Backwave is the small amount of carrier leaking through the keyed stage. This

is minimized by use of a frequency multiplier following the keyed stage and class C amplifier stages biased to beyond cutoff in the key-up condition. Keying the oscillator, of course, will eliminate backwave completely. When very low backwave is necessary, chirp due to oscillator keying can be avoided by employing a keyed stage *with wave shaping and an associated circuit which turns on the oscillator instantly* when the key closes. The chirp is over then before the keyed-stage output rises appreciably. A time delay keeps the oscillator keyed on until the keyed-stage output has decayed nearly to zero. This system is called *time-sequence keying.*

FIG. 41. Types of keying circuits.

A mixer exciter will have no backwave on the carrier frequency if the mixer stage is keyed so the carrier frequency does not exist in the key-up condition.

Two typical keying circuits are shown in Fig. 41. They may be keyed directly by a hand key, by a relay, or by a radiotelegraph machine.

31. Frequency-shift Keying (FSK). This form of modulation has been widely adopted for radioteleprinter use because it produces a definite signal on both mark and space. It gives improved circuit reliability and the ability to operate with lower signal-to-noise ratios. The standard amount of frequency shift is 850 cps. The mark frequency is 425 cps above the assigned carrier frequency and the space frequency is 425 cps below it. Other values of shift, such as 600 cps to as low as 40 cps, have been used. Wave shaping is employed to limit the band width required for passing the third or fifth harmonic of the shortest mark-space cycle.

A FSK unit is commonly used to provide a keyed input signal to the exciter of a conventional c-w transmitter. These units contain a 200-kc master oscillator which

FIG. 42. Block diagram of a frequency-shift-keyer unit.

is frequency-shift-keyed by means of a reactance tube across the tuned circuit. This keyed output is mixed with the output of a crystal oscillator or other stable frequency source in the range of 1.8 to 4.3 Mc. The keyed unit output in the 2.0- to 4.5-Mc range is fed to the transmitter exciter where it is multiplied by 2, 4, or 8 to generate the carrier frequency. The amount of frequency shift in the 200-kc oscillator must be reduced by the amount of multiplication following the keyer unit to maintain 850-cps shift at all carrier frequencies. Figure 42 shows a block diagram of a typical circuit.

32. High-level Plate Modulation. This method of amplitude modulation has been very popular because it permits high over-all transmitter efficiency and there are no critical tuning adjustments. A typical class B modulator with a modulated

class C r-f power-amplifier circuit is shown in Fig. 43. The class B audio stage is called the modulator, although the modulation actually takes place in the class C power-amplifier stage. The operating conditions of the class C amplifier are chosen

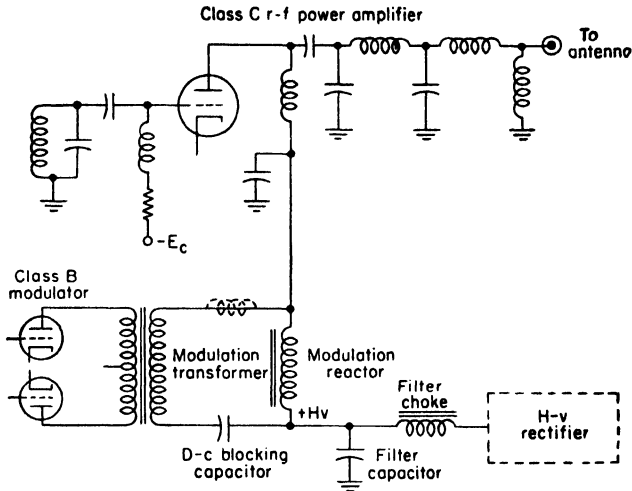



FIG. 43. Circuits for plate modulation.

so that its r-f output voltage is directly dependent upon the voltage supplied to it by the d-c power supply and the modulator (see Fig. 44). Enough r-f grid drive is required so that the r-f plate voltage is always in a saturated condition. This means that more grid drive would not have a noticeable effect on the r-f plate voltage. The fixed bias E_c is usually sufficient to cut off plate current if excitation fails. A grid resistor is used to develop enough bias so that the tube operates in the general region of twice cutoff. This allows the grid-bias voltage to vary over the modulation cycle and results in more linear modulation. Some large transmitters employ a small amount of grid-modulating voltage for the purpose of allowing the use of less driving power and for obtaining lower modulation distortion.

FIG. 44. Voltage relationships in plate modulation.

The peak audio-power-output capability of the class B modulator is equal to the unmodulated d-c plate-power input of the class C stage at 100 per cent modulation. For example, the over-all efficiency of a 10-kw transmitter PA stage may be 75 per cent, which includes plate efficiency and r-f output network losses. This requires a d-c plate input of 10/0.75, or 13.33, kw.

The peak modulator-power output must then also be 13.33 kw. The average power is half of this, or 6.67 kw, for a sine-wave modulating voltage.

The average power depends upon the audio wave shape and varies from perhaps 1 kw for speech or music audio input to 13.3 kw with square-wave audio input in this example. The average modulation level is quite low (20 to 30 per cent) in some classes of service, such as broadcasting. A saving can be made in the size of the modulation transformer in these cases. It must still be capable of the full peak power on a short-time basis (a few minutes), but the low average power results in less heating, and hence allows a size reduction.

The sizes of the modulation transformer, reactor, and blocking capacitor are determined largely by the low a-f response desired. The size of these components can be kept to a minimum by choosing their values to form a high-pass filter[1] at low audio frequencies as shown in Fig. 45*a*. This causes higher audio current in the high-voltage filter capacitor which necessitates an increase in its size.

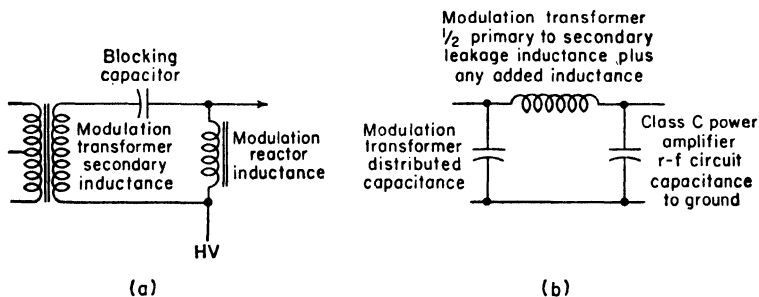



FIG. 45. High-pass and low-pass filters in modulator circuit.

Finding the optimum design with respect to size, cost, and frequency response generally requires numerous trial computations. The high-frequency response is determined by the primary-to-secondary leakage inductance of the modulation transformer along with its distributed capacitance and the effective capacitance of the r-f stage. These elements can be designed to form a low-pass filter as shown in Fig. 45*b*. Usually it is desirable to keep the modulation-transformer leakage inductance at a minimum and add the necessary inductance with a separate small inductor. It is necessary to have very low leakage inductance between halves of the primary winding to minimize distortion due to the modulator switching transients.[2] These techniques are generally used in large transmitters which have an a-f range of 50 to 7,500 cps or greater.

High-fidelity transmitters below the 1-kw power level usually use a modulation transformer designed to carry the d-c plate current of the class C stage in its secondary. This increases the size of the modulation transformer, but this is more economical since the modulation reactor and the blocking capacitor are eliminated. In voice-communication transmitters, this is usually done up to the 5-kw power level. The frequency response of these transmitters is about 150 to 3,500 cps.

33. Plate Modulation of Tetrode Class C Amplifiers. The screen of a tetrode r-f power amplifier must be modulated along with its plate. This is necessary to avoid flattening of the positive modulation peaks. The amount of d-c screen-voltage modulation is not critical and varies from about 50 to 75 per cent depending upon the tube type. A screen winding on the modulation transformer can be used, as shown in Fig. 46. It is essential that the screen-voltage modulation be in phase with the plate-voltage modulation. Differences in leakage inductance and distributed capaci-

[1] Lee, Reuben, "Electronic Transformer and Circuits," 2d ed., p. 192, Wiley, 1955.

[2] Sah, A. Pen-Tung, Quasi Transients in Class B Audio-frequency Push-pull Amplifiers, *Proc. IRE*, **24**, 1522, November, 1936.

tance along with different relative reactive loads on the two modulation-transformer windings can cause the voltages to get out of phase, particularly at high audio frequencies. The screen by-pass capacitor value in the class C stage is a compromise

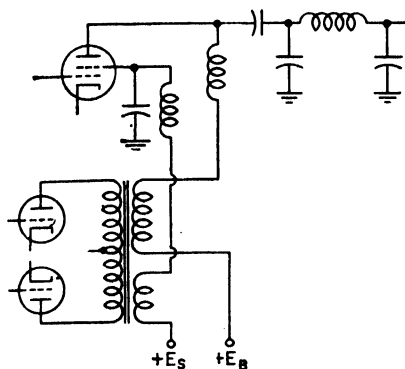



Fig. 46. Modulation of screen and plate.

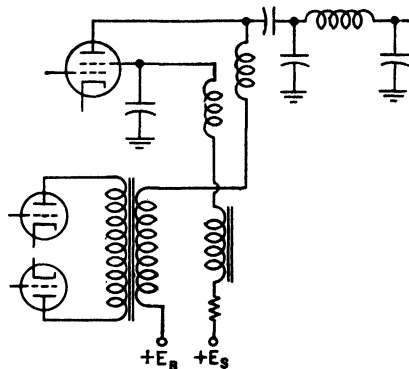



Fig. 47. Self-modulation of class C tetrode screen.

between good r-f by-passing and the shunting effect of this capacitance on the screen-modulation circuit.

The screen of most tetrodes will modulate itself if a suitable impedance is placed in series with the d-c screen supply. A resistor may be used but this requires a much higher screen-supply voltage. A reactor with 5 or 10 henrys inductance can be used

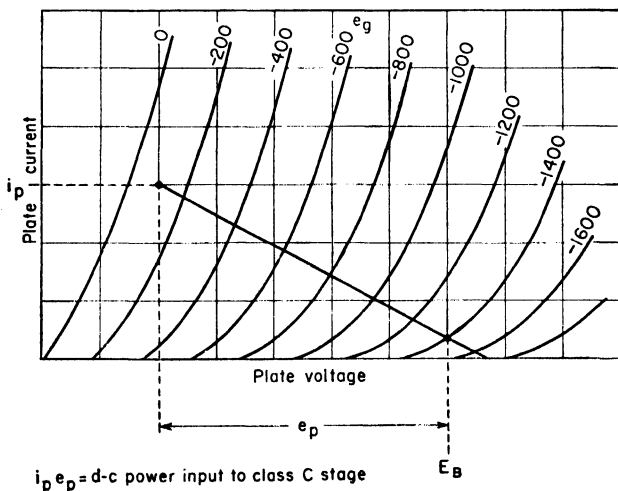



Fig. 48. Class B modulator load line.

with most of the higher-powered tetrode tubes.[1] Figure 47 shows a circuit which uses such a reactor in series with a resistor since some resistance is desirable in the screen circuit of class C tetrode amplifiers to reduce the possibility of excessive screen dissipation with variations in grid drive.

[1] Eitel McCullough, Inc., "The Care and Feeding of Power Tetrodes," Application Bulletin No. 8, pp. 11–12, 1951.

34. Class B Modulator Design. *The data sheets for most tubes suitable for modulator service list maximum tube ratings and show a typical operating condition.* It is quite easy to compute an optimum operating condition for other d-c plate voltages or power outputs, however. The modulator tube chosen must deliver a peak power output equal to the class C amplifier d-c plate-input power. This peak power is equal to the modulator-tube peak plate-voltage swing multiplied by its peak plate current as shown in Fig. 48.

The end point of the load line should be established at a point at least 5 per cent below the maximum limit to allow for variations in tubes, imperfect power-supply regulation, etc. From the end point of the load line, the modulator-tube operating conditions can be computed using the following equations.

$$\frac{2i_p}{\pi} = \text{d-c plate current per pair of tubes} \qquad (44)$$

$$\frac{\pi e_p}{4E_B} = \text{plate efficiency at maximum signal with sine wave} \qquad (45)$$

$$i_p e_p = \text{peak audio power output} \qquad (46)$$

$$\frac{i_p e_p}{2} = \text{rms sine-wave power output} \qquad (47)$$

$$4\frac{e_p}{i_p} = \text{plate-to-plate load impedance} \qquad (48)$$

$$i_p\left(\frac{E_B}{\pi} - \frac{e_p}{4}\right) = \text{plate dissipation per tube} \qquad (49)$$

With a sine-wave audio signal the modulator plate dissipation is greatest when the signal level is such that the peak plate swing is $(\pi/4)^2$ or 61.7 per cent of the modulator d-c plate voltage.

FIG. 49. Class B modulator with low-mu driver.

The choice of modulator zero-signal plate current is a compromise between distortion and zero-signal plate dissipation. A value that produces a zero-signal plate dissipation of one-half the maximum rated tube-plate dissipation is frequently used.

The d-c plate voltage is the same as that used for the class C r-f stage unless a separate modulator power supply is employed. For economy and to avoid the possibility of operating the modulator without a load, a common high-voltage power supply is generally used.

Triode, tetrode, and pentode tubes are all suitable for class B modulators. The choice of modulator tubes is often dictated by the desire to use the same tube type in the modulator as is in the r-f power amplifier. Most low-powered transmitters use beam-power tetrode tubes. In the medium-power range up to 5 kw both triodes and tetrodes are used. Triodes are found in most high-powered transmitters.

The drive for modulator tubes which are driven into grid current must have very good output-voltage regulation. The use of low-mu triodes which have low plate resistance and are coupled to the modulator grids with an impedance-stepdown driver transformer is one method (see Fig. 49). Another method uses driver tubes connected in a cathode-follower circuit, as shown in Fig. 50.

The driver transformer must have a frequency response much greater than the a-f range to be covered. This is necessary because the modulator-grid current is

rich in harmonics and these must be tightly coupled to the driver tube to realize a low driving-source impedance. The transformer phase-shift characteristics are also very important in broadcast transmitters which use audio feedback for distortion reduction.

Some modulator tubes have a grid-current characteristic with a negative slope in it, as shown in Fig. 51. A resonant circuit from grid to ground is all that is necessary for oscillation to take place when the grid voltage is in the region between A and B. The leakage inductance of the driver transformer and the input capacitance of the tube, stray capacitance, and transformer capacitance form a parallel-resonant circuit often in the range of 25 to 50 kc. This parasitic oscillation will show up in the modulator output wave (Fig. 52). Since this negative slope is due to grid emission, it varies considerably from tube to tube. Loading the grid circuit will help in marginal cases, but in others the loading required to prevent oscillation is prohibitive, and the only solution is to eliminate the driver transformer or use another tube type.

FIG. 50. Class B modulator with cathode-follower driver.

The grid-bias supply for class B modulators must have excellent d-c and dynamic regulation. The screen supply for tetrode modulators also must be well regulated.

A solution to these problems is to use modulators operating class AB_1. The plate efficiency is nearly as good with properly chosen tubes, and the grid-bias and driver-regulation requirements are eliminated. Also the driver transformer can be eliminated and either direct or capacitive coupling from driver plate to modulator grid can be used. This eliminates some phase shift which is advantageous when feedback is used. Beam-power tetrodes can be employed as class AB_1 modulators in the low-to medium-powered transmitters, and low-mu triodes (*e.g.*, mu = 4 to 5) are used in medium- and high-powered transmitters. Figure 53 shows a circuit used in a 10-kw transmitter.

FIG. 51. Modulator grid-current characteristic.

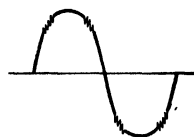



FIG. 52. Parasitic oscillations due to negative grid-current slope.

Audio feedback is usually employed in broadcast transmitters to lower distortion and improve the frequency response. Feedback voltage is taken from the modulator plates and is fed back over three stages.

The frequency-response and phase-gain characteristic of the feedback loop must be controlled for about four octaves beyond each end of the desired a-f range to permit the feedback to be effective at all audio frequencies within the required range.

35. Grid, Screen, and Suppressor Modulation. *Prior to the widespread use of high-level modulators, modulation was accomplished by grid-modulating the power* amplifier or modulating an earlier r-f stage.

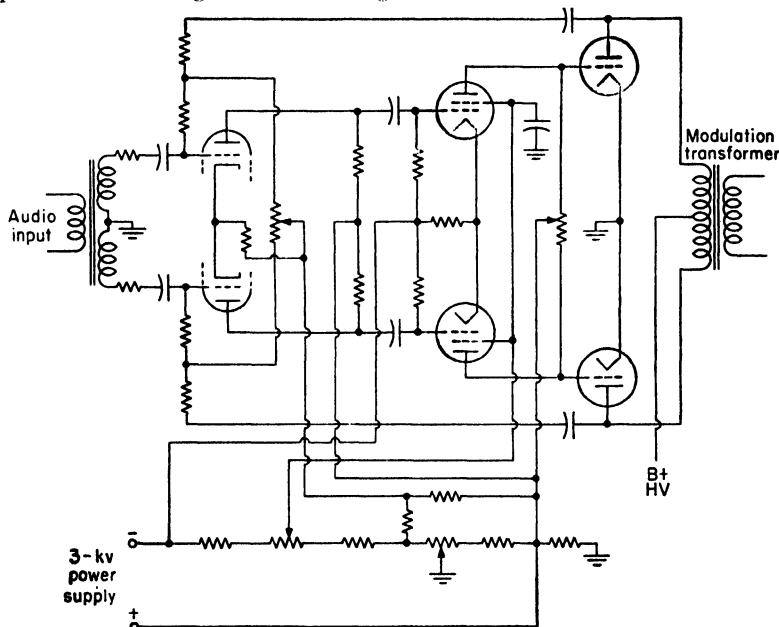



FIG. 53. Class AB_1 modulator.

Grid, screen, and suppressor modulation are all types of efficiency modulation. The audio voltage on the modulated grid varies the efficiency of the r-f stage, resulting in an a-m output signal. Figure 54 shows how the r-f output voltage varies with grid bias. The fixed bias which determines the carrier level should produce an r-f output voltage of about one-half maximum value. The efficiency of a grid-modulated r-f stage varies from zero at the trough of a 100 per cent modulated envelope to a maximum of about 70 per cent at the envelope peak. The efficiency at carrier level is one-half the maximum value, or about 35 per cent. Similar curves are obtained for screen-voltage variation and for suppressor-voltage variation. For no distortion, this curve would be a straight line up to the maximum level.

FIG. 54. Variation of r-f output voltage with grid bias.

The grid current which is modulated varies greatly with signal level and presents a varying load to the modulator tube. A heavily swamped modulator tube or a circuit with low internal impedance will reduce excessive distortion from this cause.

Because of their low efficiency and relatively high distortion, grid-modulated final r-f amplifier stages are seldom used. Exceptions are transmitters with a grid-modulated driver stage followed by a high-efficiency linear amplifier and small transmitters where efficiency or a-m power level is of secondary importance.

36. High-efficiency Linear Amplifiers. Several types of high-efficiency linear amplifiers have been developed for obtaining a relatively high efficiency at carrier level without the use of the large class B modulator necessary for high-level modulation. In general, they use two tubes. One operates near maximum efficiency at carrier level and is modulated downward on the negative modulation peaks. The second tube functions only on the positive modulation peaks.

One type still being manufactured uses the Doherty circuit[1] (see Fig. 55). Ninety-degree networks couple the grid and plate of the carrier tube to the grid and plate of the peak tube, respectively. The peak tube is biased so that it operates only on the positive modulation peaks and is cut off on the negative peaks. The 90-deg networks cause most of the r-f load to transfer from one tube to the other over the modulation cycle.

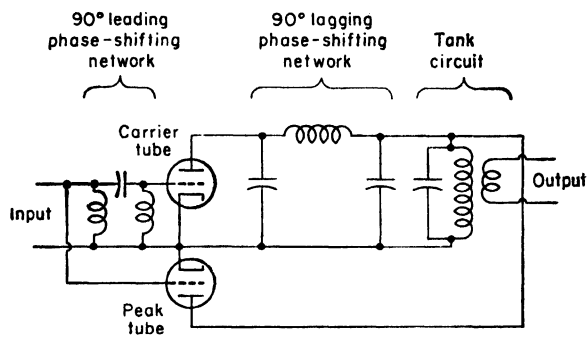



FIG. 55. Doherty amplifier.

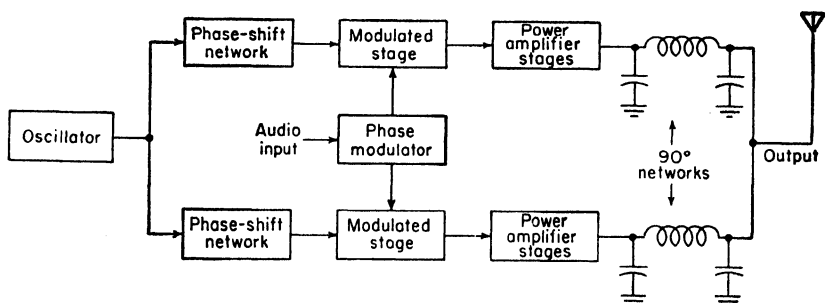



FIG. 56. Block diagram of an outphasing modulation system.

37. Outphasing Modulation. This is another type of high-efficiency a-m modulation in current use.[2] A block diagram is shown in Fig. 56. The crystal oscillator feeds two exciters which are phase-modulated with equal amounts of phase deviation but in opposite directions. Each phase-modulated signal is amplified by a separate string of identical class C amplifiers. The outputs of the two class C power amplifiers are combined using 90-deg networks. The transmitter output is the vector addition

[1] Doherty, W. H., A New High Efficiency Linear Amplifier for Modulated Waves, *Proc. IRE*, **24**, 1163, September, 1936. Witty, W. M., Modern Fifty-kilowatt Broadcast Transmitter, *Electronics*, **28**, 168, March, 1955.

[2] Chireix, H., High Power Outphasing Modulation, *Proc. IRE*, **23**, 1370, November, 1935. Starner, C. J., J. W. Lawson, and C. D. Mulford, The RCA "Ampliphase" Fifty, *Broadcast News*, **84**, 54, August, 1955.

of the outputs of the two amplifiers as shown in Fig. 57. This causes the load on each power amplifier to vary from zero (neglecting r-f network losses) to a maximum at the modulation crest. The r-f plate swing on the tubes remains at a full class C value at all times, resulting in high plate efficiency.

The Doherty amplifiers and the outphasing system both require numerous critical tuning adjustments. This is satisfactory for single-frequency operation but is the

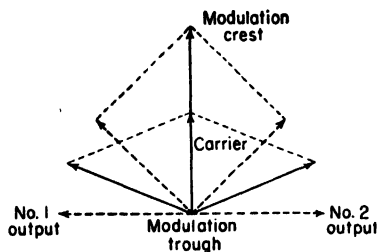



Fig. 57. Vector relationships for producing amplitude modulation.

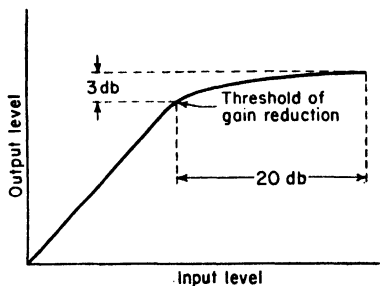



Fig. 58. Limiting-amplifier gain characteristic.

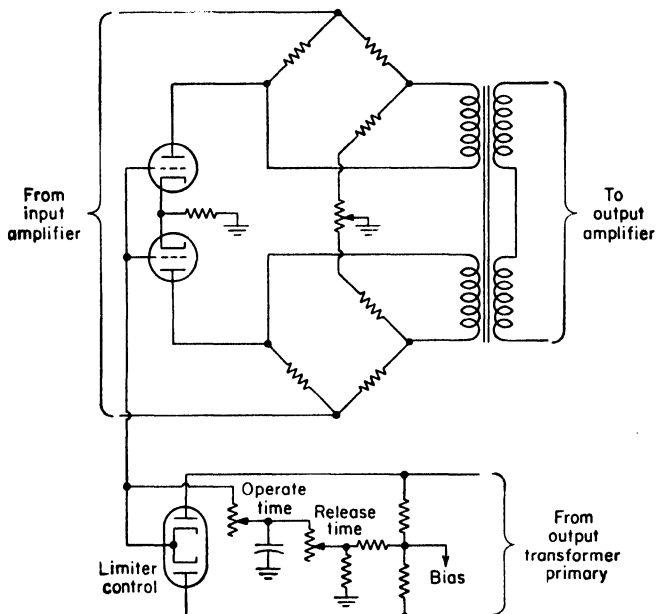



Fig. 59. Limiter control circuit.

reason these circuits are not used in equipment where the operating frequency must be changed often.

38. Audio-limiting Amplifier. The audio signal should be maintained at a value that will frequently cause 100 per cent modulation on audio peaks. This will help make full use of the transmitter-power capability. Audio-limiting amplifiers are used for this purpose. They contain a gain-reducing circuit that is operated by audio peaks that exceed a given level. Figure 58 shows a typical gain-vs.-signal-level curve.

Many circuits will accomplish this purpose. One is shown in Fig. 59. Limiting amplifiers are often built into the audio section of communications transmitters. Separate units are also used, particularly in broadcast stations. These amplifiers employ a short time constant on gain reducing and a slow time constant on gain return. These are often called the *attack* and *release* time constants. Release time constants vary from as short as 0.1 sec for speech in communications transmitters to 2 sec in some broadcast-station applications.

39. Speech Clipping. This is another method of increasing the average power in the audio signal. Instead of reducing gain, the audio peaks which exceed a given level are chopped off (Fig. 60). The vowel sounds contain much of the speech power while the consonants contain much of the intelligence. Since the vowel peaks are about four times the amplitude of the consonant sounds, it has been found that clipping off half the peaks (6 db of clipping) has only a perceptible effect on speech quality, but it raises the transmitter's average a-m side-band power by approximately 6 db.

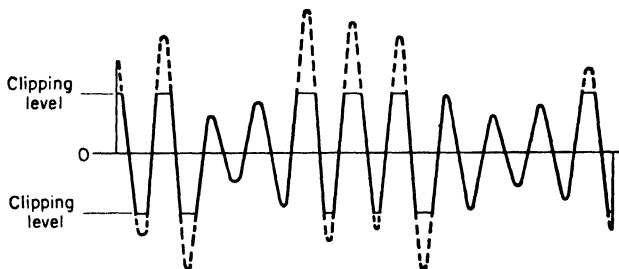



Fig. 60. Effects of speech clipping.

Up to 20 db of clipping has been used, and although the naturalness of the voice is degraded, the intelligibility remains high. A low-pass filter following the clipper removes high-frequency harmonics which are above the desired audio range.

Some of the rough sound due to the distortion generated by speech clipping can be eliminated by reducing the audio response of the lower audio frequencies where the strong vowel sounds occur. Less clipping is then required to keep the consonant sounds at full signal level. An audio limiter ahead of the speech clipper will maintain a high average level to the clipper for an additional improvement in modulation effectiveness.

The low-frequency response of the amplifiers and modulators following a speech clipper must be excellent to avoid appreciable tilting of the flat top of the clipped wave.

This type of speech clipping is not desirable in phase-modulated transmitters. An automatic peak-deviation control is more effective.[1]

Clipping audio peaks is not effective in limiting SSB envelope peaks because the audio input wave and the corresponding SSB envelope shape are different. Automatic load control (ALC), which employs a gain-reducing r-f amplifier in the exciter actuated by power-amplifier signal peaks, is more effective.

40. Frequency and Phase Modulation. Generating circuits for these types of communications transmitters are very similar. In fact, the nature of f-m and p-m signals is so similar that one can be converted to the other by applying a 6-db-per-octave slope to the a-f response. A phase modulator will produce f.m. when the audio response falls off with increasing frequency at the rate of 6 db per octave across the audio range. Most p-m circuits have a practical limit as to how many degrees the phase of the carrier can be modulated. If more deviation is required, the modulation is done at a subharmonic of the carrier. Frequency multipliers raise the modulated output to the desired carrier frequency. The degrees of phase deviation are increased

[1] Winkler, Marion R., Instantaneous Deviation Control, *Electronics*, **22**, 97, September, 1949.

by the same multiplication factor. The p-m or f-m signal is generated in the exciter, and class C r-f amplifiers raise the power to the rated transmitter carrier output.

The use of a phase modulator for generating f.m. has the advantage that a stable crystal oscillator can be used for generating the carrier. An f-m master oscillator has relatively poor frequency stability so that its use requires additional circuits for keeping the average frequency precisely on carrier frequency.

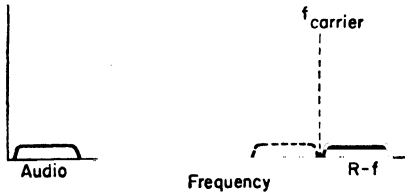



FIG. 61. SSB signal in r-f spectrum.

Frequency modulation has a signal-to-noise advantage over a.m. and SSB when the signal is strong enough to cause limiting in the receiver and when large deviation is used. This, of course, consumes more spectrum space and hence is limited to the v-h-f range and above. It is widely used in v-h-f vehicular communication because the transmitter is simpler and less expensive and consumes less power than a corresponding a-m transmitter. It is also widely used in multichannel systems in the u-h-f and microwave regions.

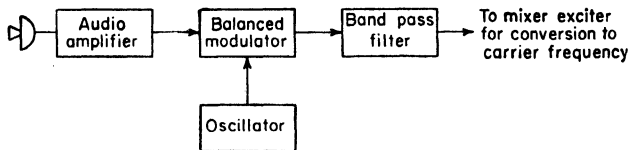
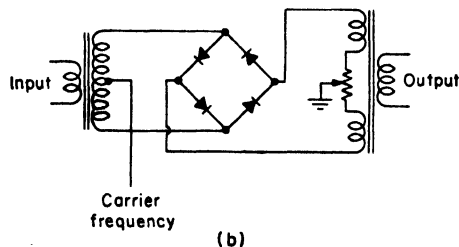
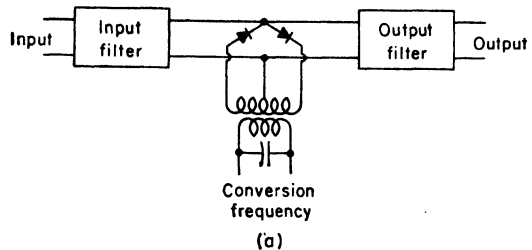



FIG. 62. Filter type of SSB generator.

41. Single-side-band Generator. The function of an SSB generator is to convert the audio signal to an r-f signal either with or without inversion. An SSB signal is one side band of an a-m signal. Figure 61 shows the frequency spectrum occupied by an SSB signal. The other side band and the carrier are substantially eliminated. The term used to express the amount that the undesired side band is reduced is *side-band suppression*. Carrier suppression may be expressed as decibels below the peak envelope power of the desired side band.

FIG. 63. Balanced-modulator circuits.

A block diagram of the filter method of SSB generation, shown in Fig. 62, is used in nearly all commercial SSB transmitters. The audio signal is mixed with an oscillator at an intermediate frequency in a balanced modulator. The output of the balanced modulator contains both side bands but the carrier is substantially eliminated depending upon the degree of balance realized in the modulator. The filter selects the desired side band and rejects the other. The output of the filter is an SSB signal at an intermediate frequency, which

is then raised to the carrier frequency by frequency conversion.[1] Several types of balanced-modulator circuits are suitable for SSB generators. A pair of semiconductor diodes can be used, as shown in Fig. 63*a*, but a diode quad, sometimes called a *varistor*, as shown in Fig. 63*b*, will have fewer undesired output products. Vacuum-tube diodes can also be used in these circuits; however, the balance tends to be affected by the cathode-to-filament capacitance and by filament voltage. A twin triode can also be used, as shown in Fig. 64. A special beam deflection tube, shown in Fig. 65, performs well as a balanced modulator and has the advantage of operating at a higher impedance level.

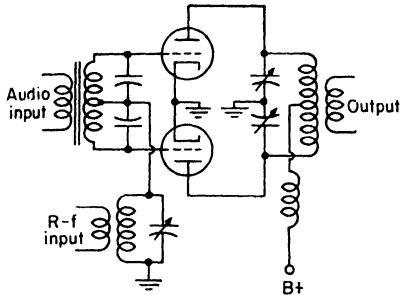



FIG. 64. Tube type of balanced modulator.

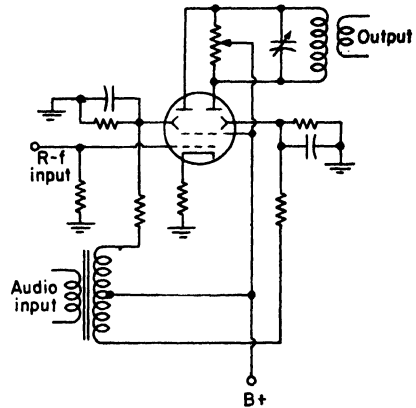



FIG. 65. Beam deflection tube used as balanced modulator.

The diode modulators generally work best at circuit impedances of about 1,000 ohms.

The filter should have a flat response across the desired speech band and very steep skirts to obtain a high degree of undesired-side-band suppression. An i-f frequency of 20 kc is used in some transmitters because satisfactory filters can then be built using high-*Q* coils and capacitors. A frequency this low requires an additional conversion step over exciters which generate the SSB at a higher i-f frequency. Very excellent crystal filters have been developed for use in the 100-kc region and have been in use for many years in transoceanic radiotelephone equipment.[2] Mechanical filters[3] in the range of 250 to 455 kc are being used extensively in voice-communication transmitters.[4] Figure 66 shows a typical response curve of a pair of mechanical filters which will select either side band at a carrier frequency of 300 kc.

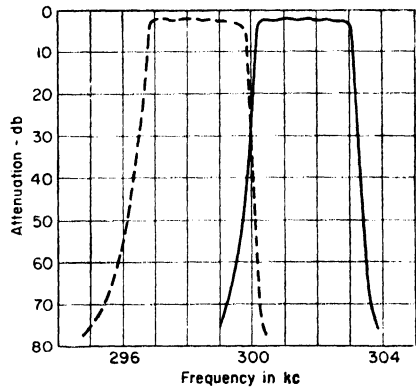



FIG. 66. Typical response of mechanical filter for SSB selection.

[1] Kerwien, A. E., Design of Modulation Equipment for Modern Single-sideband Transmitters, *Proc. IRE*, **40**, 797, July, 1952.

[2] Oswald, A. A., A Short-wave Single Sideband Radio Telephone System, *Proc. IRE*, **26**, 1431, December, 1938. Klenk, L. M., A. J. Munn, and J. Nedelka, A Multichannel Single Sideband Transmitter, *Proc. IRE*, **40**, 783, July, 1952.

[3] Doelz, M. L., and J. C. Hathaway, How to Use Mechanical I-F Filters, *Electronics*, **26**, March, 1953.

[4] Laport, E. A., and K. L. Neumann, A New Low-power Single Sideband Communication System, *RCA Rev.*, **16**, December, 1955.

The i-f frequency of the SSB generator must also be suitable for a frequency-conversion scheme that will have a minimum of undesired mixing products that may appear in the transmitter output.

Two or more SSB signals can be multiplexed together and passed through a common exciter and linear power amplifier. One method, better known as *independent side band*, uses the upper side band of one audio channel and the lower of the other. Figure 67 shows a block diagram of a typical basic circuit, and Fig. 68 shows the spectrum occupied by each channel.

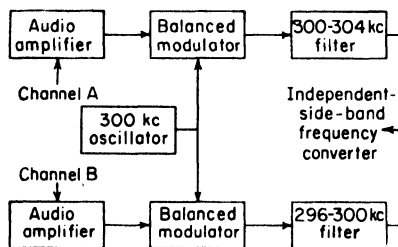



FIG. 67. Independent-side-band generator.

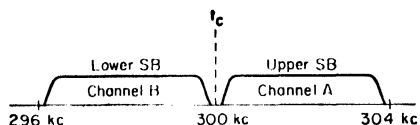



FIG. 68. R-f spectrum of independent side band.

Another method which occupies the same spectrum space but uses the upper side band for both channels is shown in Figs. 69 and 70.

In this example, a separate carrier frequency is used for each channel with 4-kc spacing. Additional channels may be added in the same manner.

Radiotelegraph channels may also be added. For example, FSK oscillators operating at the proper i-f frequencies can provide several radiotelegraph channels in addition to voice channels.

The phasing method of SSB generation[1] has been widely used in amateur radio equipment. Figure 71 shows a typical circuit. A double-balanced modulator is fed two audio and two r-f signals, each of which must be 90 deg apart in phase. When connected as shown, the phase relationship of the side bands generated in the double-balanced modulator is such that the carrier and one side band are balanced out and only the other side band appears in the output. Reversing the two audio inputs will select the other side band. The amplitude and phase of the input signals must be adjusted and maintained to a high degree of accuracy for good side-band suppression. It is very difficult to achieve 40 db of side-band suppression, which is necessary for multiplexing applications. For this reason it has been limited to a great extent to amateur service where lower values of side-band suppression are usable.

FIG. 69. SSB generation using separate carrier frequencies for each channel.

FIG. 70. Two-channel spectrum.

POWER SUPPLIES

42. Power Sources.[2] Fixed transmitters up to 1-kw power rating usually use a single-phase a-c power source. Exceptions are mobile equipment which may operate

[1] Villard, Oswald G., Jr., A High-level Single-sideband Transmitter, *Proc. IRE*, **36**, 1419, November, 1948.

[2] See also Chap. 17.

on 6 to 28 volts d.c. They may use dynamotors or vibrator power supplies to obtain the high voltages required. In the larger aircraft, 400 cps a.c. is generally used. Mobile equipment using these various types of power supplies is found in police cars and taxis, in jeeps, on ships, and in aircraft.

Above 1 kw, a three-phase a-c source is generally used. In the United States a three-phase 230-volt 60-cycle a.c. is generally available. In some countries, a three-phase 380-volt 50-cycle source is very common.

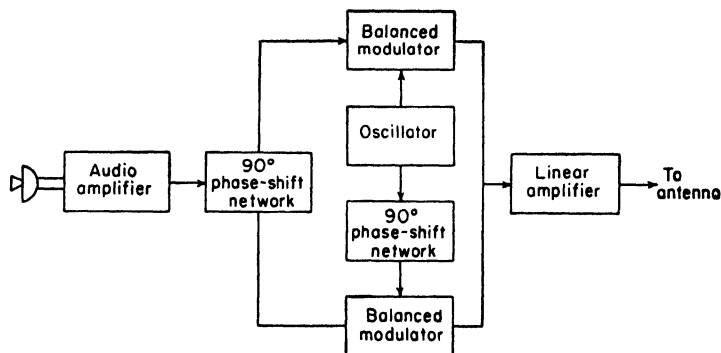



Fig. 71. Phasing type of SSB generator.

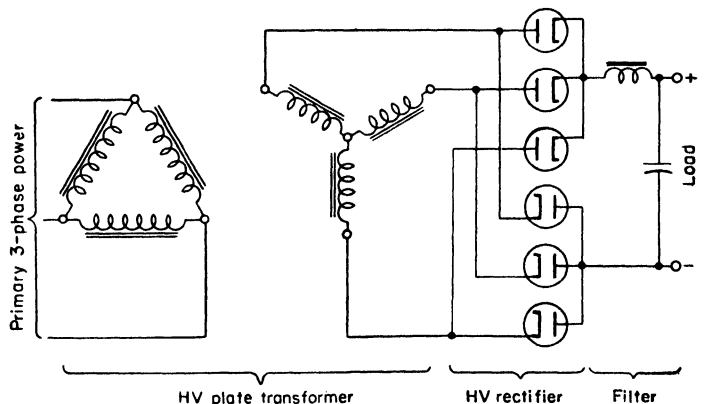



Fig. 72. Three-phase full-wave rectifier.

Three-phase full-wave rectifiers have the best plate-transformer utilization, good regulation, and a low ripple output. The ripple frequency is six times the primary a-c frequency, and it is 25 db down from the d-c plate voltage at the input to the filter. For this reason, the filter requirements are relatively small. Figure 72 shows the basic circuit of a three-phase full-wave rectifier used in a high-voltage power supply.

Other types of rectifier circuits are sometimes used to achieve certain advantages. The three-phase half-wave circuit, shown in Fig. 73, requires only three tubes. The ripple frequency is only three times the power-line frequency; so much more filtering is required. A given rectifier tube can deliver only one-half as much d-c voltage in this circuit compared with that of a full-wave rectifier. Figure 74 shows a three-

phase double-Y circuit often used to obtain up to twice the d-c output current obtainable from a three-phase full-wave bridge with a given type of rectifier tube.

It is desirable to connect the rectifier-tube filament transformers so the filament voltage on each tube is 90 deg out of phase with its plate voltage. This minimizes the

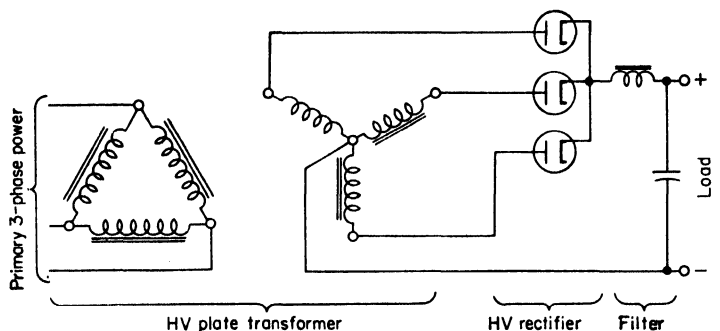



FIG. 73. Three-phase half-wave rectifier.

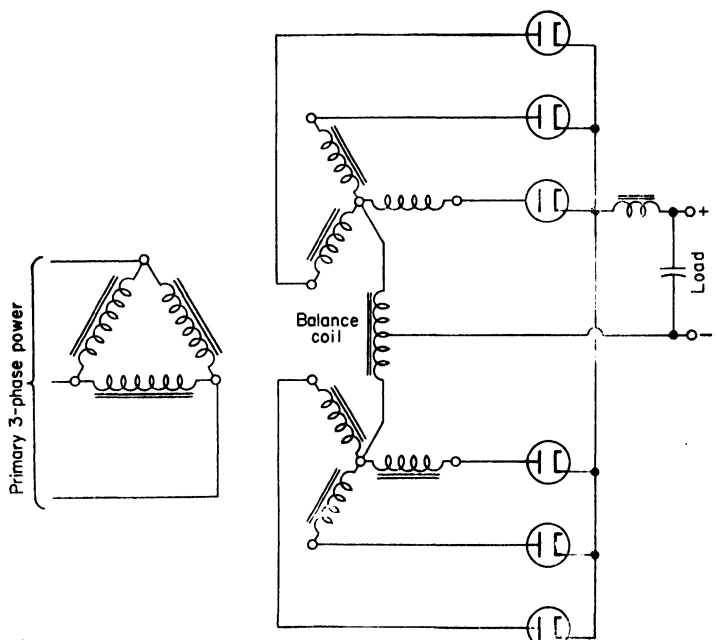



FIG. 74. Three-phase double-Y rectifier.

difference in voltage from each end of the filament to the plate. More uniform emission over the entire filament results, which increases tube life. Some of the larger rectifier tubes have increased ratings in quadrature operation. In some cases a 60-deg phase difference is more practical and nearly the full advantage is realized.

Requirements. In addition to supplying the required d-c voltage and output current, the principal power-supply requirements are adequate d-c regulation, good

dynamic regulation, and low ripple or noise output. Most high-voltage transmitter d-c power requirements are for either an "on-and-off" or a varying-load characteristic. In general, very good d-c regulation is required. High-voltage filters are nearly always used between the rectifier and the load. Generally, this is essential to reduce the output ripple and noise to very low values and avoids modulation of the signals in the amplifiers which receive power from these supplies. The filter chokes inherently place a relatively high impedance between the rectifier and the load so that it is necessary to have fairly large values of capacitance in the output side of the filter. The filter capacitor must be large enough to supply the rapid variations in load current which are impeded by the filter choke. This is particularly necessary when supplying power for class B modulators and class B linear r-f amplifier stages.

43. Rectifier Operation. Small low-voltage power supplies which have a relatively constant load can use vacuum rectifiers. Where better regulation is required, it is necessary to use gas-type rectifiers. They have a low voltage drop which remains fairly constant in the 10- to 15-volt region for all values of current up to their maximum rated values. In general, they have high current capabilities and the low voltage drops result in excellent rectifier efficiency. Mercury-vapor rectifier tubes are the most common gas-type rectifier used. In general, they have very long life when operated within their rated operating conditions. The operating temperature is generally limited to that which will permit the coolest part of the tube to be in the temperature range of 20 to 60°C. The condensed mercury-vapor temperature range is usually listed on the tube data sheets.

When frequent mercury-rectifier-tube failures are encountered, it is often found that they are being operated below their rated temperature. Operating them either too hot or too cool changes the mercury-vapor pressure in the tube from the design value and reduces the peak inverse-voltage capability. This makes them more susceptible to an arc-back. The mercury will condense on the coolest spot in the tube, which should be near the base and within the proper temperature range. There will be some temperature rise in the tube during operation because of filament heat and plate dissipation, which will raise the temperature about 10 to 15°C above the ambient air after about 30 min of operation. Very little temperature rise of the tube envelope occurs during the 30- to 60-sec filament-warmup time, however. This means that the free-air ambient temperature should not be less than the minimum temperature requirement for starting, and it should be 10 to 15°C less than the maximum rated condensed-mercury temperature. A small blast of air near the base of the tube will reduce this 10 to 15°C if operation near the maximum tube temperature is necessary.

The most critical period of rectifier-tube operation is in starting, because a transient voltage peak may occur when the plate transformer is switched on. It is very important in high-voltage circuits for the condensed-mercury-vapor temperature to be correct at this time. Some equipment is provided with heaters to raise the tube temperature when operation at low ambient air temperatures is required.

Equipment which may be subject to wide ambient-temperature variations, such as military or remotely operated stations, uses inert-gas tubes such as the 3B28 and 4B32 in place of the 866A and 872A mercury-vapor equivalents. This type of tube can be operated in ambient temperatures of -75 to $+90$°C. It is pointed out, however, that the life of these tubes is generally shorter and a rough estimate is that they have about one-third the life of their mercury-vapor equivalents. Apparently the gas tends to "clean up," which lowers the gas pressure and reduces the peak inverse-voltage capability.

Vacuum rectifier tubes are used in very-high-voltage circuits because of their higher peak inverse-voltage capability. Also, they are not temperature-sensitive. They have a large voltage drop, however, which results in lower efficiency and poorer regulation, but in the usual high-voltage low-current applications this is acceptable. Typical applications are in high-powered pulsed r-f amplifiers and high-powered klystron amplifiers.

Metallic rectifiers, such as selenium, copper oxide, and magnesium sulfide types,

have been frequently used in power supplies delivering less than 100 volts for energizing relays, etc. They have had limited application in high-voltage transmitter supplies, although one 50-kw broadcast transmitter uses selenium rectifiers and another uses germanium rectifiers exclusively.

Germanium and silicon junction diode rectifiers have a very low voltage drop, which results in excellent regulation and efficiency. Developments in the field of semiconductor junction rectifiers are very rapid, and it is expected that they will find very wide application. The top ambient-temperature range for germanium rectifiers is in the range of +55 to +85°C, while the silicon types can operate at temperatures up to +150°C and even higher. They are filling many applications where the advantages of small size, low weight, and excellent efficiency are important. Rectifier

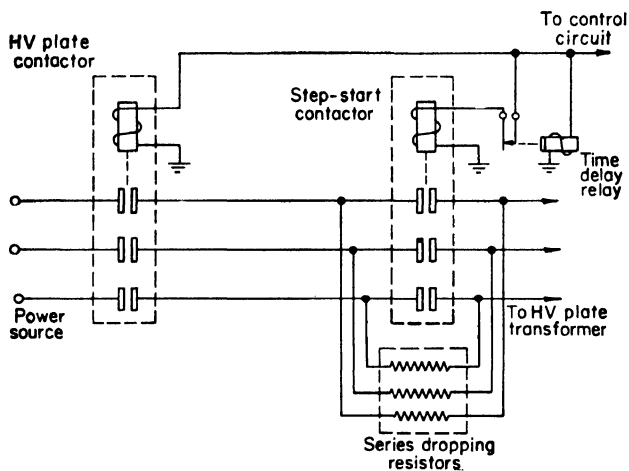



FIG. 75. Step-start circuit.

stacks using elements in parallel or in series to obtain higher current or voltage usually have matched rectifier elements. Germanium and silicon rectifiers have very small size and thermal capacity for their ratings, and hence protection against large transients is often necessary.

44. Power-supply Transients. Transient voltages and currents which far exceed the steady-state values are frequently encountered in high-voltage power supplies. These transients tend to be greater when transformers with high-grade steel operate near their maximum flux density. The primary inrush current transient is determined largely by the residual flux in the core, the primary coil inductance in air, and the point on the primary voltage wave at the instant of closure. For this reason, the transient may vary greatly from one switch closure to the next. Peaks up to ten to twenty times the maximum rated line current may be encountered.

Plate-transformer secondary-voltage-transient peaks may be twice the normal secondary voltage and can be observed on an oscilloscope. An examination of repeated starts on both single-phase full-wave circuits and three-phase full-wave circuits showed occasional peaks of twice the normal a-c voltage. The principal significance of these peaks is that they can cause rectifier tubes to arc back if this peak exceeds their inverse-voltage capability. For this reason, it is desirable to operate rectifier tubes in circuits where the normal peak inverse voltage does not exceed one-half the rated peak inverse voltage. When this is not possible, *step-start* circuits are sometimes used. One type starts the transformer with resistors in series with the primary leads, as shown in Fig. 75. After a short time delay (usually less than 1 sec), these resistors are shorted, which brings the rectifier up to full voltage operation.

Greatly reduced primary current transients and a smaller secondary voltage transient result.

A third kind of power-supply starting transient is caused by the charging current required to bring the voltage up on the filter capacitors. This current is determined largely by the capacitance involved, the inductance of the chokes, and the value of the steady-state d-c supply voltage. This transient can cause excessive current in the rectifier, and to limit it some large high-voltage power supplies start with a resistor in series with the filter capacitor (Fig. 76). The capacitor charges through this resistor, which is shorted out after a time delay.

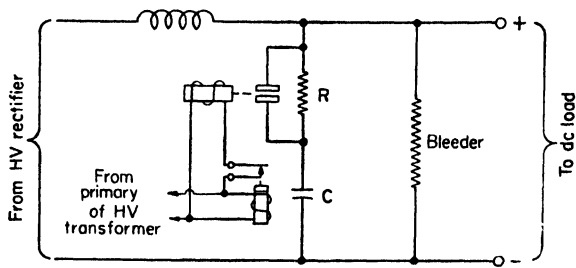



FIG. 76. Filter-capacitor charging-current-surge limiter.

CONTROL CIRCUITS

Small transmitters, such as those used in mobile service, may have very simple control circuits. The "on" switch applies power to the heaters of all cathode-type tubes. A push-to-talk button on the microphone controls a relay, which in turn applies plate voltage and puts the transmitter on the air. This keeps the power drain at a minimum and reduces wear on the high-voltage generator, which is usually a vibrator or dynamotor. Some transmitters use quick-heating filament-type tubes in the power stages, which also are operated by the push-to-talk relay for further reduction of stand-by power drain. Larger mobile transmitters will have some of the control-circuit functions described in the next paragraphs.

The number of control-circuit functions depends to a large extent upon the size and complexity of the transmitter. Push buttons are usually used to initiate certain control functions. Relays then complete the function in the proper manner. These circuits provide necessary protection to tubes and components against improper sequencing and malfunction. Also, interlock provisions are made to minimize the possibility of personnel coming in contact with dangerous high voltages. These three functions of circuit control, equipment protection, and personnel protection are interrelated and are all integrated into the transmitter control system. The function and purpose of various parts of control circuits in common use are as follows.

45. Filament On-Off Switch. The filament-control push button or switch initiates a sequence of functions that applies power to all filaments, and starts the cooling system and time-delay circuits that make the transmitter ready for the application of plate power. When operated to the "off" position it shuts the transmitter down.

46. Filament Heater Control. The filament "on" switch operates a relay which energizes the filaments of all heater-type tubes and the filaments of the high-voltage rectifier tubes.

47. Blower Control. One or more blowers are used to cool the tubes and to provide general transmitter cooling and are usually started by the filament "on" circuit. The life and reliability of many components are greatly dependent upon temperature; so proper cooling is essential. Many high-power tubes require the application of cooling before filament heating power is applied. Air interlocks prevent application of power to tubes which require cooling when the blower is not operating for any reason. Some air interlocks have a small vane in the blower-outlet air stream to close a switch when the air velocity is adequate. Another type uses a switch actuated by the air pressure developed in a plenum chamber by the blower. A "blower off" delay circuit is often provided in large transmitters to cool tubes and components after

all power is removed from them when shutting down. The blower keeps running until a time delay of a few minutes has expired. Then the blower is automatically stopped.

48. Power-amplifier Filament. Separate relays may control power to the filament of large transmitting-type tubes. When filament cooling is required for the tube, the filament relays are interlocked with the blower air switches.

49. Time Delay. Mercury-vapor rectifier tubes require a warmup time of 30 sec to 1 min before the application of plate voltage. Cathode-heater-type tubes may require a heating time of 30 sec. A time-delay circuit which is energized when filament power is applied assures the proper time delay to avoid tube damage. At the end of the delay period, a contact in series with the high-voltage primary-control circuits is closed, which allows application of plate voltages. Motor-operated time-delay units use a gear-reduction train to operate a switch after a certain number of motor revolutions. Pneumatic, dashpot, and mercury time-delay relays depend upon the slow flow of air, oil, or mercury through a small orifice. Thermal time-delay units use a heater to cause bending of a bimetallic strip. Each has its particular characteristics and field of usefulness. Some reset almost instantly, and others, like the thermal type, reset very slowly. An ideal type for tube protection would remain operative for a second or two after power removal and then only reset at the same rate that the tube filaments cool. This achieves the shortest delay in getting the transmitters on the air following a power interruption.

50. Bias-supply Control. It is essential that bias voltage be applied to the various power-amplifier stages before plate or screen voltage. In some transmitters the bias-supply plate transformer is energized as soon as the time delay has operated. If semiconductor bias-supply rectifiers are used, bias may come on with the application of filament power. In other transmitters, the bias supply is not energized until the plate "on" switch is energized. An interlock is nearly always provided. It may be an extra contact on the bias plate relay, but a relay operated by the bias-supply output voltage will protect against bias-supply failure also.

51. Low-voltage Plate-supply Control. The plate voltage on the smaller tubes in the exciter and audio-amplifier stages is sometimes applied automatically as soon as the vacuum-type rectifier tubes warm up. If mercury-vapor rectifier tubes are used, the plate circuit cannot be energized until the time-delay period has expired. These supplies are generally large enough to warrant switching with the high-voltage supplies. Provisions are usually made so that these low- and intermediate-voltage power supplies can be operated without energizing the high-voltage supply. This is a useful feature for tuning and trouble shooting.

52. High-voltage Plate-supply Control. The transmitter control circuits are sequenced and interlocked so that everything else must be on and functioning before the high-voltage plate transformer can be energized. A single exception is the application of screen voltage to some power tetrodes. Screen dissipation in many of them will be excessive if screen voltage is applied without plate voltage. An interlock relay actuated by d-c plate voltage prevents this. Relays in small- and medium-powered transmitters, and remotely operated circuit breakers in high-powered transmitters, energize the plate transformer.

53. High-voltage Step-start Circuit. Very high current and voltage transients often occur when power is instantly applied to a high-voltage plate transformer. Primary-line-current surges may be up to ten or twenty times normal line current, and secondary voltages up to twice normal are common. An excellent step-start circuit places resistors in series with the transformer primary when the primary power circuit is closed. The resistors limit the transients to relatively small values. After a short time delay (less than 1 sec), the resistors are shorted, and full voltage is applied to the plate transformer primary.

54. Sequence Start. In many transmitters, particularly those suitable for remote control, each part of the transmitter is energized in the proper sequence automatically.

Initiation of the transmitter start function may be by a single push button or a remotely operated relay. This function may be combined with the "plate on" push button, which closes a holding relay. This energizes the filament control circuit and all circuits are energized in order, and finally the high-voltage plate power comes on.

55. Overload Relays. Medium- and high-power tubes are nearly always protected from excessive plate current by overload relays in plate or cathode circuits. These relays remove the high-voltage primary power if the plate or cathode current exceeds a predetermined value. The relay coils are shunted with a low-value resistor to avoid the effects of the reactance of the relay coil to a-f current components. The more elaborate overload relays have a flag which indicates that it has operated. The value of tripping current can be adjusted on some relays directly, and on others the shunt resistor is adjustable. Overload relays are quite effective in class C amplifiers where the plate current rises considerably when the plate circuit is out of tune.

56. Overload Recycling. Many overloads that occur when the transmitter is in normal operation will clear themselves as soon as the high-voltage d.c. is removed. A switching transient or lightning surge may cause an arc-over, for example. For this reason, most large transmitters in installations where off time must be kept at a minimum use a recycling circuit to cause the transmitter to come back on once or twice before staying off. Some transmitters will try once at full power and, if the overload persists, will go off and try again on a low-power position before shutting down. Automatic-reset provisions are usually made so that, if only one overload occurs which clears itself, the counting circuit returns to normal after several seconds of delay.

57. Primary A-c Circuit Protection. Fuses and circuit breakers are used to protect the primary circuits and some of the larger components, such as transformers and motors. The major a-c branch circuits are protected, but small transformers may be protected as a group because they have such high internal impedance that an overload would not cause enough increase in primary current to blow a fuse. Switch-type circuit breakers are common in large transmitters, because they can be used as a switch as well as for circuit protection.

58. Personnel Protection. High voltage is very dangerous to human life; therefore, precautions are made to prevent accidental contact with it. Various safety standards have been set by the different transmitter manufacturers. Protection is nearly always provided against access to voltages above 250 volts. Some rather elaborate protection schemes have been devised. Nearly any scheme can be cheated, and sometimes this is necessary when servicing the equipment. The operator or service technician interested in staying alive does not take chances.

Door interlock switches will remove all high voltage in an enclosure when the door is open or an access panel is removed. These switches are connected either in series with the plate-transformer primary circuit or in series with the transformer-control relay coil.

High-voltage shorting switches ground high-voltage circuits when a door is opened. This is an added protection used in circuits with over 1,000 volts on them because the primary door interlock can fail. Also, it assures that the high-voltage filter capacitors are discharged. A cautious operator will always use a grounding stick in addition before touching anything in a high-voltage circuit.

59. Typical Transmitter Control Circuit. A control circuit typical of those used in 5- to 20-kw transmitters is shown in Fig. 77. The requirements of each transmitter are different, and more or fewer functions may be necessary. In general, the control-circuit complexity is proportional to the power rating of the transmitter. This control circuit is separated from the primary power circuits and is fused or protected by a separate circuit breaker. Small transmitters frequently have their control circuits mixed in with the primary circuits. This may save some relays and is acceptable in low-power circuits, since the primary fuses or other protections have low enough ratings to protect the control-circuit components also.

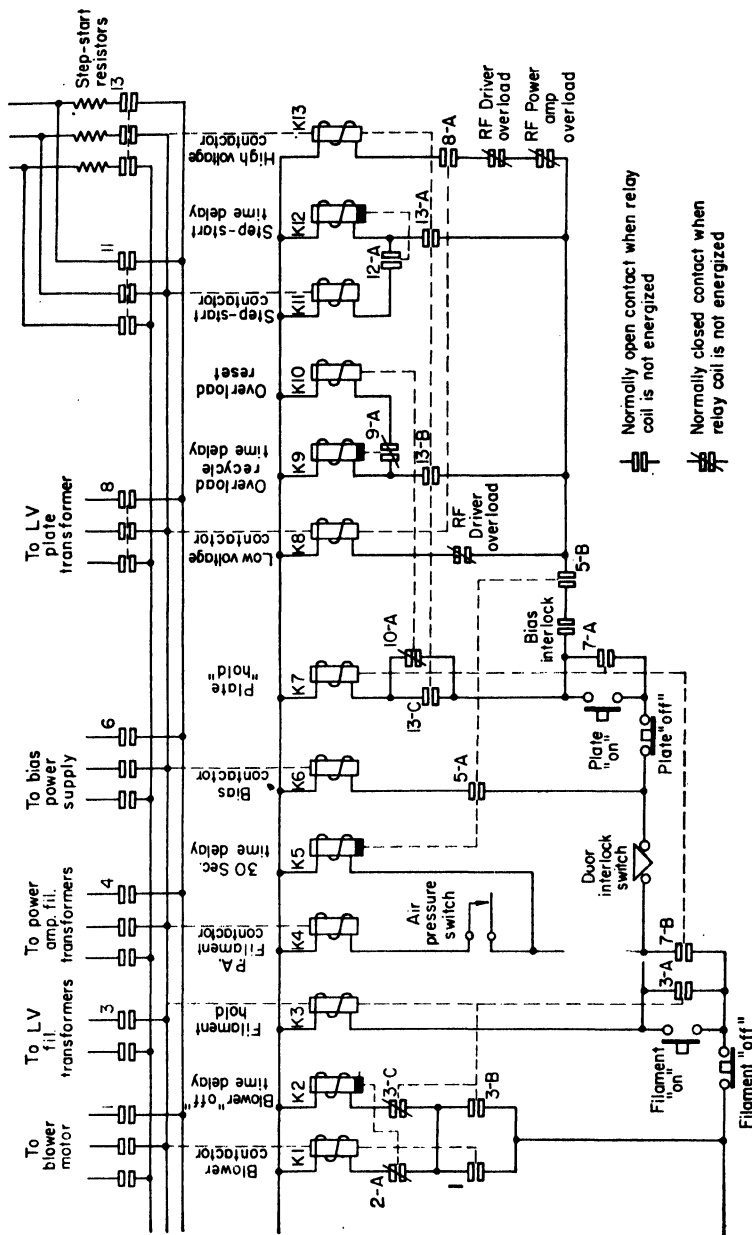



FIG. 77. Transmitter power-control circuit.

The function of each relay in Fig. 77 is shown, and dashed lines aid in associating the contacts with the proper relay coil. Before starting the transmitter, all primary power disconnect switches and circuit breakers must be closed. Pressing the filament-on button ("fil on") closes the circuit through the filament-hold relay K3. One of its contacts, K3-A, parallels the "fil on" contacts and interlocks the relay so it remains energized when the push button is released. (Note: All four of the filament and plate "on and off" push buttons are of the momentary-contact type.) This relay also may energize the filaments of all low-power tubes and the high-voltage rectifier tubes. Another contact, K3-B, causes the blower contactor, K1, to be energized, which in turn starts the blower. This blower supplies cooling air to the power-amplifier tubes as well as general transmitter cooling. When sufficient air pressure is developed, the "air pressure switch" closes and energizes the "PA fil" contactor. This interlock function assures the presence of cooling air when the power-amplifier filaments are on. The "30 sec time delay" relay K5 was also energized when the "fil on" button was pressed. Its contacts remain open for 30 sec to allow the rectifier filaments in the bias, low-voltage, and high-voltage power supplies to reach operating temperature before plate voltage can be applied to them. At the end of the 30-sec time-delay period, K5-A closes and energizes the "bias plate" contactor, K6, providing the "door interlock switch" is closed. Actually, there may be several door interlock switches in series, and all must be closed before any bias or high-voltage plate power can be applied. The primary function of door interlock switches is personnel protection, but the cooling-air system may not function properly if either a door or an access panel is open. A bias interlock relay whose coil is in the bias supply d-c output circuit becomes energized as soon as the bias supply is operating and closes the "bias interlock" contact. This prevents application of plate voltage without bias being present, as a protection to the amplifier tubes. The sequence of functions initiated by pressing the "fil on" button is now completed, and the transmitter will remain at this stage of operation until another sequence is initiated.

Pressing the "plate on" push button initiates the steps which bring the transmitter into full operation. The "plate on" switch energizes the "plate hold" relay K7 through the contacts K10-A which are closed. It holds itself closed by means of its contact K7-A when the push button is released. The "low voltage" contactor, K8, also operates, providing the 30-sec time delay has elapsed and the "bias interlock" contact is closed. The "driver overload" relay contact will be closed. Its purpose is to open the low-voltage plate-voltage-supply primary circuit if driver plate current becomes excessive. When the "low voltage" relay operates, its contact K8-A energizes the "high voltage plate" contactor K13, which applies primary power to the high-voltage plate transformer through the "step-start resistors." Closure of K13-A energizes the "step-start time delay" relay which delays closing of the "step-start" contactor by a fraction of a second. The transmitter is now in full operating condition.

The auxiliary contact K13-B on the "H.V. plate" contactor energizes the "overload recycle time delay" relay, and through its normally closed contact, K-9A, the "overload reset" relay, K10, will be energized. This opens K10-A. As long as K10-A is open, an overload in either the r-f driver or power amplifier will cause all low-voltage and high-voltage plate power to be removed. The "plate on" button must be pressed again to apply plate power.

After a period of 5 sec, the "overload recycle time delay" relay will open K9-A. This deenergizes K10, and K10-A closes. Now when an overload causes plate power to be removed, it will be reapplied immediately. If the overload has cleared itself, the plate power will remain on. If not, the second opening of an overload relay contact within 5 sec will cause plate power to remain off because K7 will be deenergized. This is called a "two-shot" overload circuit because it takes two successive overloads within a 5-sec period to shut off plate power.

Plate power may be removed by pressing the "plate off" push button. Filament power may be removed by pressing the "fil off" push button. The blower will continue to run because its contactor is kept energized by the "blower off time delay," K2. When its time delay has elapsed, K2-A opens and releases the "blower contactor," K1. The transmitter is then completely shut down.

This control circuit has a sequence-start provision. Pressing the "plate on" button operates the "plate hold" relay, which has a contact, K7-B, across the "fil on" push button. The "plate hold" relay holds itself on by means of K7-A. The entire transmitter will come on with each step in the proper sequence and can be shut down by pressing the "fil off" button, which turns off all plate power. The blower continues to run until the blower time delay has expired as before.

CARRIER-FREQUENCY CHANGING

60. Frequency-changing Needs. Some communications services require nearly instantaneous shift in carrier frequency. One example would be to answer a calling aircraft on the frequency it may be monitoring. In some applications a few seconds' delay is acceptable, and in others a minute's delay will cause little inconvenience. In still other circuits, there is little advantage in having quick frequency-shift provisions at all, and a frequency change time of even 5 or 10 min is acceptable. Some h-f communication stations and short-wave broadcast stations change frequency on a fixed schedule. Also, there are those which change to a new frequency when the propagation conditions require it for communicating to some fixed point; and it is frequently necessary to use different frequencies when communicating to stations located at various distances from the transmitter.

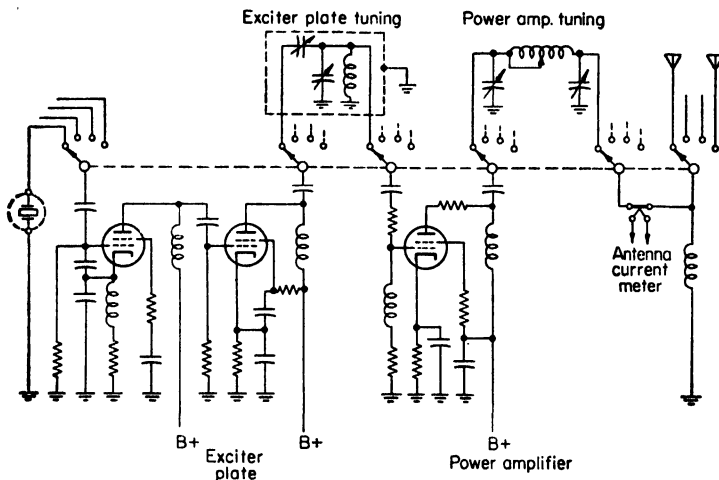



FIG. 78. Simplified schematic of band-switched transmitter.

61. Channel-switching Transmitters. An instantaneous frequency change is usually accomplished by some means of switching. Some of the h-f transmitters in the 50- to 100-watt class, which are set on the operating table, switch tuned circuits by a ganged wafer switch operated by a control knob on the front panel. A selection of frequencies is provided, and plug-in components allow any frequency to be set up on each channel. Figure 78 shows a simplified schematic of such a transmitter. Any channel may be selected simply by rotating the band switch when all circuits are properly tuned.

This same principle is used in 200- to 1,000-watt transmitters, except that relays or some type of remotely operated switches are used. Two- and four-channel transmitters are most common.

62. Multiple R-f Bay Transmitters. A very popular method of quickly changing frequency in the 2- to 3-kw power size is accomplished by switching complete r-f units. Several units are connected to a common power supply. A common modulator is used for amplitude modulation. The power supply generally has enough capacity to operate two or three r-f units simultaneously. As many as 12 separate r-f units are used in some installations with a common power supply. The crystal ovens and filaments of the heater-type tubes are operated continuously. When a channel is selected for transmission, the keying and modulator circuits are switched to the r-f unit which is tuned to the selected frequency. The thoriated-tungsten filaments, the blower, and the low-voltage circuits are energized by additional relay switching. The

channel-change time depends upon the switching circuitry, but usually it takes less than a second or two. Figure 79 shows a simplified diagram of such a switching method.

This type of communication transmitter is commonly located several miles from the operator's position and is remotely controlled over telephone lines. Telephone-type switching is used to turn the transmitter on and off, and to select the channel and type of emission. A second control function is keying or push-to-talk carrier control.

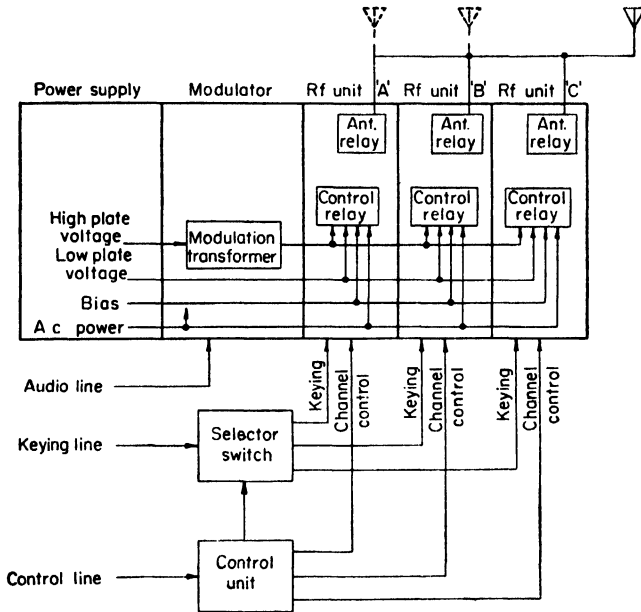



FIG. 79. Block diagram of a switched multibay transmitter.

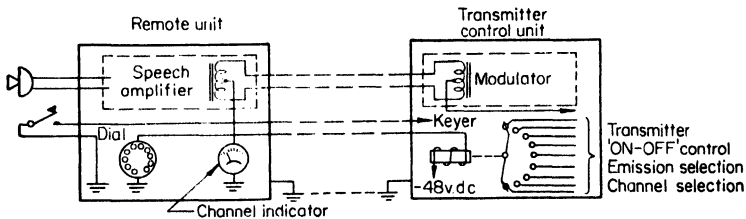



FIG. 80. Simplified four-wire remote-control system.

A third remote connection is required to carry audio for voice modulation. Sometimes a remote indication of the on-off condition of the transmitter and the channel selected is furnished. Two telephone pairs can perform all these functions. Figure 80 shows the basic circuits used to accomplish this. When the remote distance is several miles or more, these functions all usually are carried on one nonmetallic telephone pair, *i.e.*, one having a repeater or transformer in it so that d.c. cannot be transmitted. One tone (60 cps) is used for dial pulsing and a second (3,500 cps) for keying. Filters at the transmitter separate these tones and rectify them to create the

d-c voltages required by the control circuits. Figure 81 shows a block diagram of such a system.

63. Autotune[1] Transmitters. A second basic method of changing frequencies is one which mechanically tunes each control to a preset position set up for each channel. One method uses an Autotune system in which the tuning conditions for up to 10 channels can be preset. Figure 82 illustrates the basic principle. When changing to a new frequency, the tuning control rotates counterclockwise to its stop or home position. The cam drum then rotates to the position which will allow the pawl corresponding to the selected channel to drop. The tuning-control shaft on which the stop drum

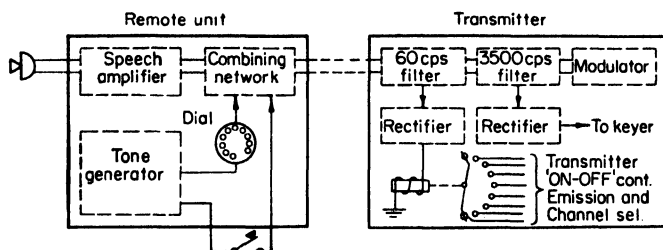



FIG. 81. Simplified two-wire remote-control system.

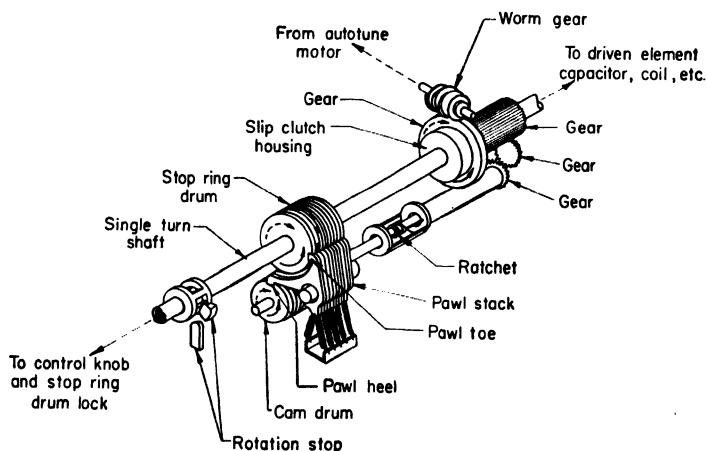



FIG. 82. Autotune principle.

is mounted rotates clockwise until the pawl drops into the notch which stops rotation. The clutch which couples this mechanism to the driving motor then slips until the motor is stopped automatically. Figure 83 shows the basic circuit of a locally controlled Autotune system.

Telephone-dial control can also be employed for remote operation of the Autotune. The frequency-change time generally is on the order of 3 to 30 sec, depending on the type of Autotune used. Figure 84 shows a simplified schematic of a dial-control circuit for remotely operating an Autotune system.

64. Servomechanism Tuning. In another method used in several medium- and high-powered transmitters for frequency changing a servomechanism controlled by a potentiometer is provided for each tuned element. A bank of preset potentiometers

[1] Registered trade-mark of Collins Radio Co.

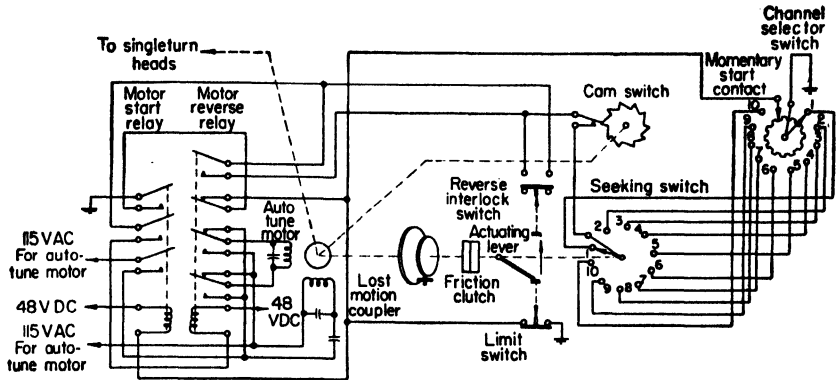



FIG. 83. Electrical details of the Autotune system.

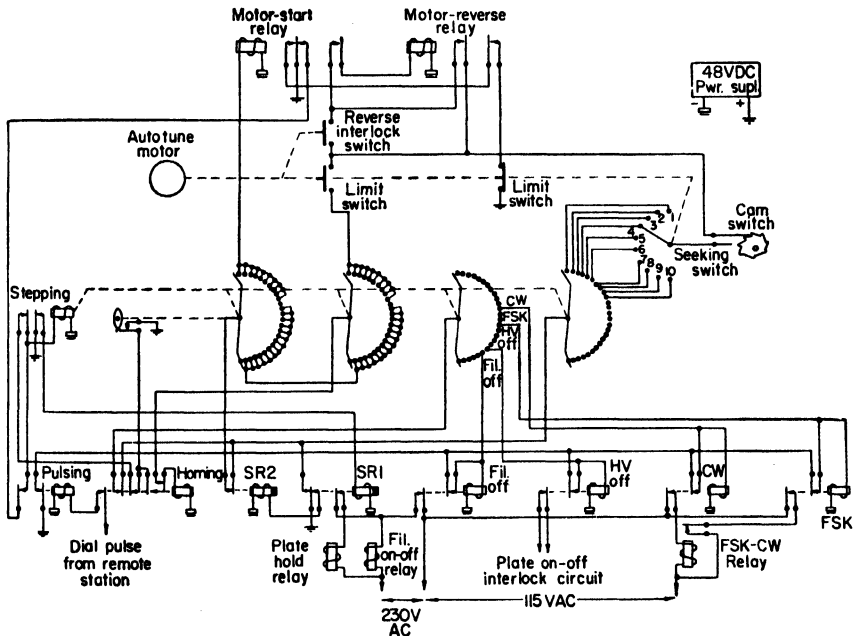



FIG. 84. Control of remote Autotune system.

may be provided, and the group which has been adjusted for a particular carrier frequency can be selected by a switch. Figure 85 shows a simplified schematic of a servomechanism with the switching provided for several control potentiometers. This circuit is duplicated for each servomechanism, and the switching for all potentiometers is ganged together.

65. Autopositioners.[1] Mechanical positioning of circuits which are tracked to provide a linear tuning curve can be accomplished by means of Autopositioners. These mechanisms are able to position a shaft accurately to any of an integral number of

[1] Registered trade-mark of Collins Radio Co.

shaft positions per revolution. For example, an accurate stop position may be obtained for 8, 10, 12, or 20 equally spaced stops per revolution. This type of positioning has been used in the v-h-f and u-h-f transmitting equipment which allows any selection from hundreds of channelized operating frequencies.

66. Automatic Tuning and Loading. Servo-tuned transmitters can be made completely automatic by obtaining a control signal directly from the tuned circuits.

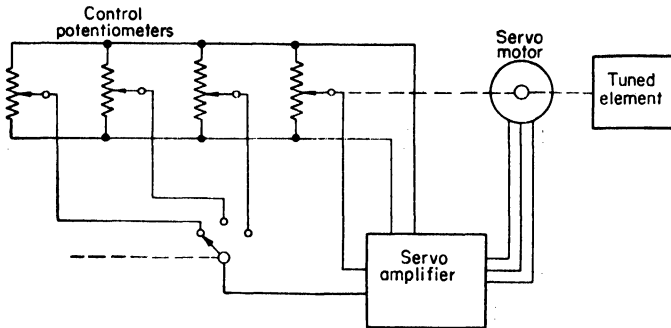



FIG. 85. Simplified schematic of a servomechanism.

R-f phase detectors compare the phase of the r-f plate voltage on a tube with the grid voltage. A d-c signal is obtained when the plate voltage is not 180 deg out of phase with grid voltage and the d-c polarity indicates to which side of resonance the plate circuit is tuned. This information can be used to operate the servo and cause it to tune the circuit to resonance. A typical phase-detector circuit is shown in Fig. 86. The range of pull-in is less than an octave; so wide-range transmitters use some means of coarse tuning to approximate frequency. One method uses a wide-band frequency detector (see Fig. 87). Its output is positive or negative d.c. which may be combined with the phase-detector output to obtain both coarse tuning and accurate fine tuning by just sensing the r-f signal.[1] Combined detector voltages are shown in Fig. 88. This type of circuit is capable of a tuning accuracy of ±5 deg, which is usually closer to exact resonance than an average operator will tune a stage by observing plate-current dip. The phase detector provides a very sensitive indication of resonance.

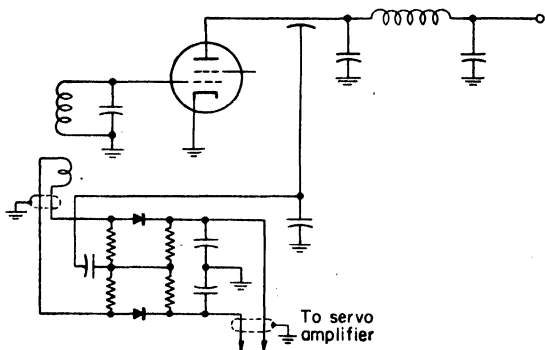



FIG. 86. Phase-detector circuit.

Automatic loading of a power amplifier may be accomplished by adjusting loading for the correct value of plate current in a class C amplifier. Usually a fixed ratio of d-c plate current to d-c plate voltage is used.

Class B linear amplifiers may be loaded for the correct ratio of r-f plate swing to r-f grid swing.

The elements of the output network must be designed and ganged so they can be

[1] DeLong, V., Automatic Tuning for High-power Transmitter, *Electronics*, **29**, 134, July, 1956.

adjusted completely by one tuning and one loading control unless other information is also provided. The coil may be band-switched from information received from the frequency-setting control on the exciter, for example. Figure 89 shows two ways an output network may be ganged. It is essential that rotation of any tuning control in a given direction will always produce the same effect. For example, a clockwise rotation of the loading control should always increase loading. If it ever changes effect under any conditions of load or mistuning, the servo system will lose directional sense.

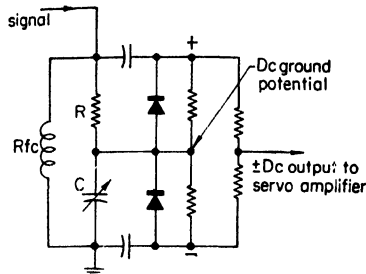



FIG. 87. Broad-band frequency-detector circuit.

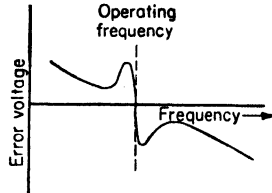



FIG. 88. Combined output of phase and broad-band frequency detector.

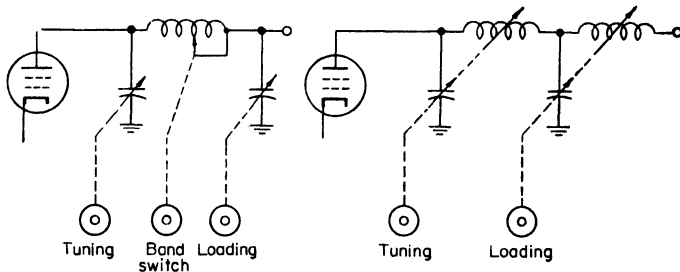



FIG. 89. Ganged output networks for automatic tuning and loading.

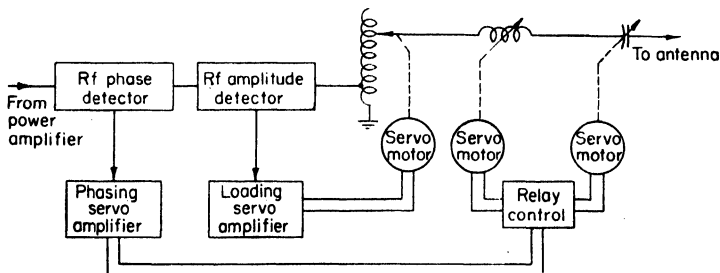



FIG. 90. Automatically tuned antenna coupler.

67. Automatically Tuned Antenna Coupling. A simplified schematic of an automatically tuned antenna coupler is shown in Fig. 90. These have become widely used in aircraft installations. The antenna coupler is located where the antenna enters the fuselage. It matches the antenna to a coaxial cable which connects the coupler unit to the transmitter. Two servomechanisms are required. The r-f detectors are located at the 50-ohm coupler input. One detector compares the phase

of the line current with the line voltage. Its output is positive or negative d.c. depending upon whether the r-f current phase leads or lags the phase of the r-f voltage. This causes the one servo to run in the proper direction to produce a pure resistive input impedance. The second detector compares the relative amplitude of the input current and input voltage. It causes its servo to run until the ratio of voltage to current equals the characteristic impedance of the transmission line.

When this type of antenna coupler is used, the transmitter output network can be designed to match the power amplifier to the characteristic impedance of the transmission line with only one tuning servo. Careful ganging and tracking of the network elements are essential to maintain constant loading over a wide frequency range. This system is most commonly used in air-borne installations and in some mobile ground stations.

ANTENNA-FEED METHODS

The antennas used at transmitting and receiving stations contribute greatly to the effective communicating power of a system. A piece of wire long enough to load the transmitter and working against a poor ground can easily require several times as much transmitter power as a simple dipole of the correct height. A directional antenna that concentrates the radiated power in the desired direction and at the optimum vertical angle of radiation will give further improvements of 3 to 20 db, providing a stronger signal at the receiver or permitting use of less transmitter power.

Practically all good antennas use a transmission line to carry the r-f power from the transmitter to the antenna. This may be a few feet of small coaxial cable in the case of a v-h-f radio installation in a taxicab to ½ mile of 600-ohm open-wire line feeding a rhombic antenna. Some power is always lost in a transmission line, a loss which must be balanced against the optimum antenna location and a convenient transmitter location.

68. Transmission Lines. Since the use of transmission line is becoming almost universal, the transmitters are being designed to feed lines that are most common to each class of service. For example, m-f broadcast stations use vertical antennas working against a ground system; so they are most conveniently fed with an unbalanced line. Unbalanced-line impedances of 50 and 75 ohms are quite standard, but unbalanced lines using four, six, or more wires and having an impedance of about 250 ohms are used also in some installations to reduce transmission-line cost. Broadcast transmitters therefore are designed for feeding these impedances.

In the h-f range, the 600-ohm open-wire line is most common for medium and high power. It is an efficient low-cost type of line that is easy to construct. Since it is a balanced line, it adapts itself well to feeding balanced antennas which are most common in the h-f range.

Since the radiation losses and discontinuities of open-wire lines become excessive in the v-h-f and u-h-f regions, coaxial line is generally used in this range. Wave guides are used from the u-h-f region on up in frequency.

69. PA Output Network Considerations. The trend in transmitter power-amplifier design has been away from push-pull and toward wider use of single-ended circuits. Inductive coupling also is being used less in h-f transmitters which must operate over a wide range of frequencies because of the problem of adjusting the inductances and the coupling between them. Output networks which use simple pi and T sections provide the greatest amount of harmonic attenuation with the smallest number of elements and the smallest number of tuning controls.

Some means of matching the power amplifier to the transmission line must be provided. This can be done in the transmitter output network. A 600-ohm matching network is shown in Fig. 91.

When L_{2A} and L_{2B} are separate coils with no mutual coupling, the balance section is one type of Alford network.[1] Typical values of reactance for L_{2A} and L_{2B} would be 400 ohms, but they can vary over a wide range. L_{2A} and L_{2B} must be equal.

[1] Alford, A., Matching Networks, U.S. Patent 2165086.

The reactance of C_3 is one-half the coil reactance. When mutual coupling exists between the coils, the reactance of C_3 decreases and would become zero if unity coupling existed. This circuit has also been referred to as an autotransformer with its leakage inductance tuned out by C_3. The input impedance of this balanced network contains an inductive component so that C_2 must have more capacitance than would be required for matching a resistance load of 150 ohms, which is the equivalent parallel resistive component at this point with a balanced 600-ohm resistive output load.

FIG. 91. Output network with unbalanced-to-balanced impedance transformation.

A high coefficient of coupling between the coils reduces the input shunt inductive reactance and also makes the network more broad-band. Baluns of this type are built as separate units for a 4:1 ratio of impedance transformation in the low-power range.

There are many advantages in using coaxial line within the transmitter building. The impedance is ideal for inserting additional harmonic suppression filters. This line eliminates r-f radiation near the transmitter, which is a great help in keeping r.f. out of the input circuits, control circuits, and power lines. In many cases, antenna switching can be done much more conveniently with a 50-ohm coax line switch than in a 600-ohm open-wire line.

FIG. 92. Dipole fed with shortened balun.

70. Shortened-line Balun. Many types of h-f antennas can be fed directly with a coaxial transmission line.[1] A dipole fed with a shortened balun for converting from an unbalanced line to a balanced antenna is shown in Fig. 92. The length l can be reduced from $\lambda/4$ to about $\lambda/10$ by tuning it to resonance with a capacitor. This coupling method has a band width of about 5 per cent.

71. Coil Balun. The band width of this type of balun can be increased by winding the shorted transmission-line section in the form of a coil as shown in Fig. 93. This presents a higher shunt reactance across the output over a wider frequency range. This coil balun has the advantage of being more compact which often is useful. It has the same impedance on both input and output terminals. The only difference is the unbalance-to-balance conversion.

72. Impedance-matching-line Balun. An impedance-matching balun is shown in Fig. 94. It uses the principle of the L network combined with a balun. The inductance, which determines the balun length, and the value of capacitance can be computed for the desired impedance transformation. This simple device is very useful, but its band width reduces with increasing values of impedance transformation.

73. Broad-band Impedance-matching-line Balun. A more broad-band impedance-matching balun employs several sections of transmission line with stepped values of impedance as shown in Fig. 95. This type of balun is useful over nearly a 4:1 fre-

[1] The balun transmission-line couplers discussed in this section are developments of the Collins Radio Company and have been described in its publications.

quency range.[1] The length l of each section is about $\lambda/4$ at the arithmetic mean of the highest and lowest frequency to be covered. This would be $\lambda/4$ at 12 Mc for covering the range of 6 to 18 Mc. The impedance of the transmission-line sections increases in geometric steps.

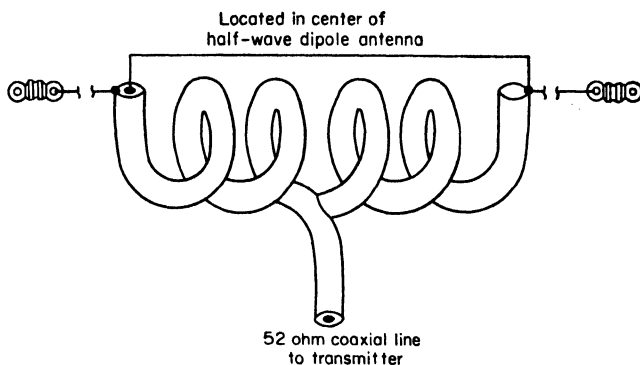



FIG. 93. Coil-type balun.

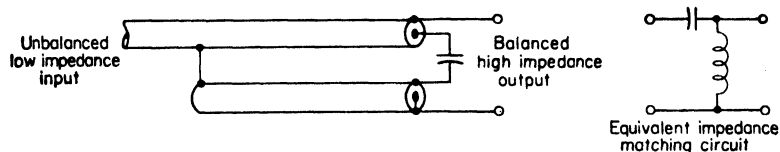



FIG. 94. Impedance-matching balun and equivalent circuit.

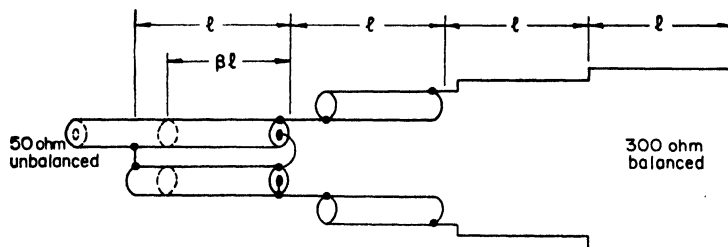



FIG. 95. Broad-band stepped-line balun.

74. Broad-band Impedance-matching Coil Balun. Another type of impedance-transforming balun uses coiled sections of transmission line as shown in Fig. 96. The parallel transmission lines may be coiled in either solenoid or spiral coil fashion. At one end the lines are connected in parallel, and at the other end they are connected in series. This results in a 4:1 impedance transformation. The coiled transmission lines pass balanced currents in the usual manner, but the coil inductance presents an impedance to unbalanced currents. For this reason, it can be rather successfully used with unbalanced as well as balanced input.

75. Broad-band 50- to 600-ohm Transmission-line Coupler. Figure 97 shows a broad-band unit for coupling a 50-ohm coaxial line to a balanced 600-ohm open-wire

[1] Fubini, Eugene G., and Peter J. Sutro, A Wideband Transformer from an Unbalanced to Balanced Line, *Proc. IRE*, **35**, 1153, October, 1947.

line. Units of this type have been built that can cover the entire 2- to 30-Mc range at power levels of a few kilowatts. It uses three coiled transmission-line sections to produce a 9:1 impedance transformation. A coil balun is used for a more perfect unbalance-to-balance conversion. The capacitors cancel some of the shunt inductance of the coil balun and the impedance-transforming coils at the low end of the frequency range. In practice, these are shunted with r-f chokes to keep the entire system grounded.

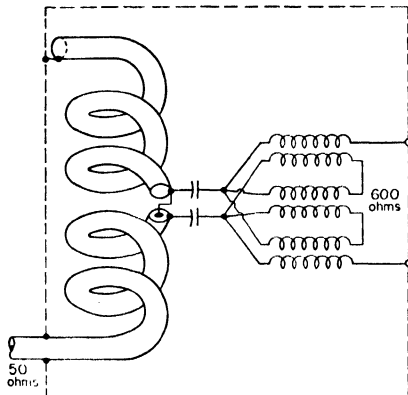
FIG. 96. Impedance-transforming balun.

This type of coupler is useful in coupling a transmitter with a 50-ohm output impedance to broad-band antenna systems, such as rhombics.

TRANSMITTER PERFORMANCE AND METHODS OF MEASUREMENT

76. Standards. The radio-station licensing authority, such as the FCC in the United States, establishes certain performance requirements. A few requirements, such as frequency tolerance, are established by international treaty. In addition to the basic requirements of frequency tolerance, power output, types of emission, band width, and carrier-frequency assignments, the FCC has published Standards of Good Engineering Practice.

The Institute of Radio Engineers has a standing committee on radio transmitters which publishes definitions of terms and methods of measurements. These are generally accepted by the industry and form a common understanding of the terms used in describing transmitter performance and the meaning of certain measurements. The Electronic Industries Association also has publications which may establish standards of performance as well as methods of measurement.

FIG. 97. Broad-band 50- to 600-ohm transmission-line coupler.

Transmitter manufacturers generally publish performance specifications on their equipment. Usually these are minimum performance figures which the manufacturer will guarantee. In other cases they are nominal performance ratings. The following paragraphs will discuss some of the more general transmitter-performance requirements and how they are measured.

77. Power Output. Power output is generally considered to be that power which the transmitter delivers at its output terminals. It always refers to average power over an r-f cycle. When the term "peak output power" is used, it means the average power over one r-f cycle at the crest of the output envelope.

The power-output rating may be given in several ways. It may be a guaranteed minimum or it may be a nominal rating. A nominal rating may mean that the power output may deviate either way from nominal, depending upon operating conditions or operating frequency. Also, it could mean the licensed rating of a broadcast transmitter which may be called upon to deliver enough extra power to overcome transmission-line and antenna-matching losses.

When a transmitter is designed to be operated on any frequency within a wide range of frequencies, there is some frequency that is least favorable to the transmitter. It may be low gain in some stage or higher tank-coil heating or any of a number of other factors. For this reason, the transmitter usually is tested for power output at this frequency and several others also. This frequency is usually at the maximum output frequency, but it may occur on the low end of the highest-frequency band or at some other frequency.

The transmitter may deliver power to certain load impedances more efficiently than to others. The power output is measured when feeding the nominal load impedance such as 52 ohms unbalanced or 600 ohms balanced with a unity standing-wave ratio, unless otherwise specified. There are numerous methods of measuring power. One basic standard is the calorimetric method. The r-f power output is dissipated in a resistive load which is liquid-cooled. The power is computed from the rate of flow of the cooling liquid and its temperature rise. When the cooling liquid is water, the power output in watts is

$$\text{Power} = 264Q\,\Delta t$$

where Q is water flow in U.S. gallons per minute and Δt is the difference between the inlet and outlet temperature in degrees centigrade. This method has the advantage that the exact impedance characteristics of the load need not be known. It can be used at any r-f frequency. The test setup must keep the heat loss or absorption from other sources, such as by conduction or radiation, to an acceptably low value.

R-f power in the low-frequency range is often measured by using a calibrated r-f ammeter in series with a known load impedance. The power is I^2R, where R is the equivalent series-resistance component of the load.

R-f loads are available for terminating coaxial lines in their characteristic impedance. Some include a meter calibrated in power. These instruments are usually r-f voltmeters calibrated in power since the load resistance is fixed.

Directional wattmeters are available for much of the r-f range. They are designed to be used in coaxial or wave-guide transmission lines and measure forward and reflected power. The real power is the difference of these readings.

There are also several substitution methods of measuring r-f power. These use a-c power to heat the r-f load the same amount as the r-f power heated it. Instruments are available for measuring the a-c power accurately, which determines the r-f power.

Transmitters which produce a type of emission that uses a carrier (*e.g.*, a.m., p.m., or f.m.) are rated in unmodulated carrier power output. The power output of an a-m transmitter is 150 per cent of carrier power when 100 per cent sine-wave modulation is applied. The peak power output is four times carrier power since both voltage and current are doubled at the modulation crest. Some a-m transmitters are not rated for continuous 100 per cent modulation. In many classes of service, the average modulation is quite low so the continuous modulation capability may be rated at some value such as 30, 50, or 70 per cent modulation.

Telegraph transmitters designed for on-off keying are not always rated for continuous key-down operation, since the key-down time seldom exceeds 50 per cent in normal transmission. When FSK is used, the power-output capability must be for continuous key-down output, however.

SSB transmitters are rated in peak envelope power output. A standard test signal of two equal-amplitude frequencies usually is used. The average power output, which can be measured by the calorimetric method, is one-half the peak envelope power, if there is no distortion present. The error due to distortion usually is neglected, since the average power measurement measures signal power quite accurately even though this may differ slightly from one-half peak envelope power. Peak-reading vacuum-tube r-f voltmeters usually have a fairly long time constant so they can be used across a known load resistance to measure peak envelope power directly. The time constant of the detector in directional wattmeters usually is very short; so it measures the average value of the current and voltage over the SSB envelope which

produces a reading of just 0.9 of the actual rms output power. For example, a 900-watt directional wattmeter reading for a two-frequency test signal indicates 1,000 watts of average power and 2,000 watts of peak envelope power. When one input frequency is removed and the other doubled in amplitude, the power output will be 2,000 watts, which will read correctly on the directional wattmeter.

The level of intermodulation distortion is very important in transmitters which multiplex voice channels. For this reason, the transmitter peak envelope power capability is sometimes established at the power level which produces a specified signal-to-distortion ratio such as 35 db.

The average power output of an SSB transmitter with voice modulation is quite low. The size, weight, and cost of the transmitter power-supply and output network components can be reduced by taking advantage of this fact. Therefore, many

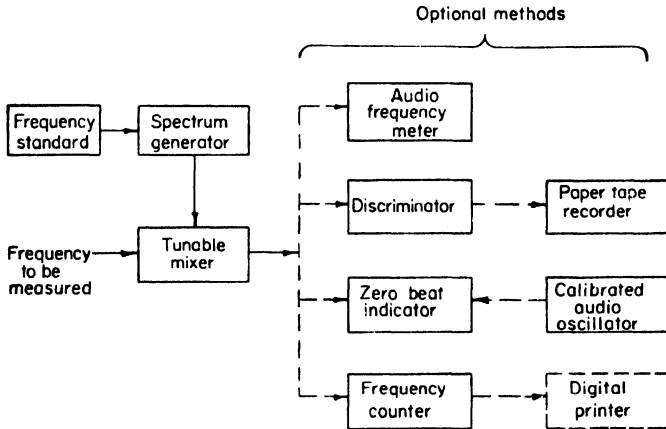



FIG. 98. Frequency-measuring apparatus.

SSB transmitters are rated for continuous operation with a two-frequency test signal but not for continuous single-tone output. Transmitters which will be used at maximum power for FSK or time-division multiplex must be rated for continuous power output, however.

78. Carrier-frequency Stability. Licensing authorities establish a tolerance within which the carrier frequency must be maintained. For example, m-f broadcast transmitters in the United States must be kept within ±20 cps of their assigned carrier frequency. A per cent tolerance such as 0.005 per cent is usually specified for transmitters in the h-f range and above.

Most transmitter specifications include the equipment performance capability with respect to frequency stability. This term is used to specify the ability of the transmitter to maintain a constant frequency. Such factors as variations in the power-source voltage, ambient temperature, relative humidity, or certain tuning conditions may affect carrier frequency. The maximum effect on frequency for specified amounts of variation is usually specified. In addition, aging may slowly change the carrier frequency. For this reason, frequency stability sometimes is given as having certain limits for a stated period of time, such as for a day or a month.

Frequency drift refers to the changes in average frequency during a specified interval of time. For example, this may refer to the change in frequency during the first few minutes of operation, or it may refer to aging effects over a long period of time.

79. Methods of Frequency Measurement. In one method, the radio wave is mixed with a standard frequency to obtain an output frequency which can be measured accurately and added to or subtracted from the standard frequency to obtain the frequency measurement. Figure 98 shows a block diagram of this type of apparatus.

The standard frequency is derived from a crystal oscillator usually operating at 100 or 1,000 kc. The spectrum generator employs frequency dividers, multipliers, and mixers to provide an output which contains a frequency at every 10-kc point across the spectrum of interest. These spectrum points have the same frequency stability as the standard oscillator. Its frequency can be checked against WWV by beating a spectrum point against a signal received from WWV. The crystal oscillator can be trimmed to zero beat.

A tunable mixer, which can be a receiver, selects the spectrum points nearest the frequency to be measured and mixes them, producing an output in the range of 0 to 5,000 cps if 10-kc spectrum points are used. This frequency can be measured and recorded by any of the four alternative methods shown. It must be determined whether the standard-frequency spectrum point is above or below the frequency of measurement, however. A button usually is provided on the crystal oscillator unit to change its frequency slightly in a known direction. Observing whether the beat frequency raises or lowers provides a means of establishing whether the beat frequency should be added to or subtracted from the standard-frequency spectrum point. The tunable mixer or receiver must be calibrated accurately enough to determine which 10-kc spectrum point is being used.

A paper-tape recorder is very useful for monitoring the frequency changes that occur during frequency-stability tests.

Another method uses a frequency counter directly. These instruments count the r-f cycles over a precise interval of time and display the resulting count. If the interval is 1 sec, the output reading will be directly in cycles per second. These instruments contain an internal standard oscillator which determines the accuracy of the time interval. An external standard-frequency source also may be used. Since the frequency range of counters is limited, some instruments contain a spectrum generator for converting the frequency to be measured down to the range of the counter. Instruments of this type are used for measurement of frequencies from audio to the u-h-f range. These instruments count for a set interval of time, display the count, and repeat this cycle continuously, if desired.

80. Audio-frequency Response. The a-f response of broadcast and voice-communication transmitters frequently is specified to be within a specified number of decibels over a given a-f range from the 1,000-cps response. For example, the response shall not vary more than ± 1 db between 50 and 7,500 cps from the 1,000-cps reference level. Another method merely requires the response to be within 2 db, for example, over the stated frequency range.

Communication transmitters usually limit the a-f response range. The attenuation of frequencies below some frequency in the range of 150 to 400 cps is used to permit a saving in the audio and modulator sections. This eliminates the low voice frequencies which contain much power but contribute little to intelligibility. The size of a modulation transformer is directly related to the lowest audio frequency it must pass at full power. The phase-modulator capabilities in an f-m transmitter also are directly related to the lowest audio frequency. The high audio frequencies are attenuated to limit the channel space occupied by the transmitter. Rapid attenuation of frequencies beyond 3,500 cps generally is required in voice-communication transmitters.

The response is measured at a sufficient number of frequencies to establish a frequency-response curve. A constant audio-input level is used for each frequency. The transmitter response is measured by a suitable detector, feeding an indicating instrument. A noise- and distortion-measuring instrument usually is used, which has a meter scale calibrated directly in decibels. The frequency response may vary a little with signal level, but it usually is measured at a moderate level such as 30 per cent modulation, unless otherwise specified.

81. Distortion. The distortion introduced by the transmitter generally means the rms value of all harmonics as a per cent of the fundamental test frequency. It may vary from less than 1 per cent for an a-m or f-m broadcast transmitter to 10 per cent

for a mobile voice-communication transmitter. The conditions under which distortion is measured are generally stated. A single test, using a 400- or 1,000-cps test frequency at 95 per cent modulation or at maximum deviation, may suffice. A complete test on a broadcast transmitter may require enough measurements to plot a curve of distortion vs. frequency over the entire a-f range. Distortion is usually greatest at maximum signal level, but sometimes data are taken also at lower levels.

A distortion meter generally is used for performing the distortion measurement. A suitable detector is used to demodulate the transmitter output signal. The instrument is adjusted to give the proper reference-level reading for the test tone. It is then adjusted to remove the test tone, leaving only the harmonics. Their level is then read directly in per cent distortion on the meter.

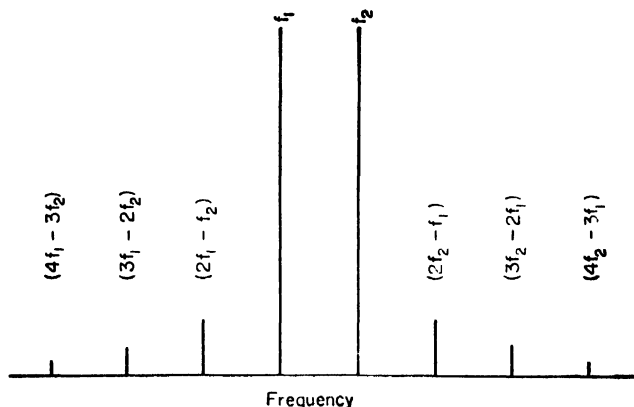



FIG. 99. SSB distortion products.

82. Signal-to-distortion Ratio. The principal type of distortion generated in SSB transmitters is intermodulation distortion. Signal-to-distortion ratio is the ratio of the amplitude of one frequency of two equal-amplitude test frequencies to a third-order distortion product. This ratio usually is expressed in decibels. Figure 99 shows the relationship of the distortion products to the two test frequencies f_1 and f_2. The inner pair of distortion products is called third-order products. They are spaced from f_1 and f_2 by the same amount f_1 and f_2 are spaced. The other products are fifth order, seventh order, etc., and all have the same spacing in frequency. The third-order products generally are strongest, although sometimes certain distortion-canceling conditions may exist in the transmitter which allow the fifth-order products to be greatest. A signal-to-distortion specification usually infers that all products are to be below the stated level.

The signal-to-distortion measuring instrument is a type of spectrum analyzer that has enough selectivity and frequency stability to select each frequency in the spectrum shown in Fig. 99 and measure its amplitude. A panoramic spectrum analyzer, which presents the spectrum of interest on the calibrated face of an oscilloscope tube, is an excellent measuring means, as all products are shown simultaneously.[1] The amplitude scale must be calibrated in decibels to obtain enough dynamic range, and this requires a logarithmic amplifier in the vertical deflection input circuit.

Signal-to-distortion specifications vary from 25 db for a single-channel low-powered voice-communication transmitter, where distortion is not important, to 40 db, where channels may be multiplexed and where out-of-channel emission must be kept to a low level. These distortion products are generated in the r-f power-amplifier stages,

[1] Icenbice, P. J., and H. E. Fellhauer, Linearity Testing Techniques for Sideband Equipment, *Proc. IRE*, **44**, 5917, December, 1956.

and the products of one channel may fall in another, creating an undesirable noise. This is more important than distortion of the signal itself.

83. Harmonic and Spurious R-f Output. As the number of transmitters and receivers increases, the probability of interference increases. Actually, interference is often most severe between various transmitters and receivers located in one station. Several h-f and v-h-f transmitters along with radar equipment may be operating simultaneously in one installation. A spurious output of one transmitter falling on the frequency of another signal being received can be serious. Some limits are being established for maximum spurious and harmonic output. These are being based on a fixed spurious output power limit so that higher-powered transmitters must incorporate a greater amount of spurious-signal suppression. For example, the maximum spurious-signal level expressed in decibels below the transmitter power for public-service transmitters approximately follows the formula 43 + 10 log power. The actual limits for various power ratings are given in a table, such as

	Db
Up to 3 watts	40
3–25 watts	50
25–150 watts	60
150–600 watts	70
Over 600 watts	80

The latest regulations and method of measurement should be consulted to determine the specific requirements for each case.

The output of a transmitter always contains some output at harmonics of the assigned frequency. The second harmonic is usually strongest, since the tank circuits provide greater attenuation of the next several higher-order harmonics. At still higher harmonic frequencies, the tank circuit may lose its effectiveness because of the stray capacitance across tank coils and the series inductance of capacitor elements. Also, resonances may occur at these frequencies so that harmonic output may be greater at the tenth or fifteenth harmonic than it is at some of the lower harmonics.

Other spurious frequencies depend upon the frequency-generating scheme. For example, a transmitter which uses frequency multipliers in its exciter may have an output on any harmonic of the fundamental oscillator frequency. Mixer-type exciters have many other spurious output possibilities which can be established by the chart in Fig. 8.

A signal-strength meter generally is used for making these measurements. The instrument must have adequate selectivity ahead of the first amplifier tube to prevent the generation of stronger harmonics in it than are being measured. The input level to the measuring instrument must be quite low; so a broad-band voltage divider is essential when measuring the output of a transmitter directly. A capacitance voltage divider coupled to the signal-strength meter by means of coaxial cable is usually most effective. The dummy load for the transmitter must present a constant impedance at all frequencies of measurement, if the measurements are to be meaningful.

Another method of measurement uses a signal-strength meter to measure the actual signal strength of undesired radiation. The transmitter may radiate some output directly besides that which is radiated by the antenna.

Undesired output may also be conducted out on the power-line and control-circuit connections. A separate measurement is necessary to determine the voltage of undesired outputs on these wires for a complete evaluation of all undesired transmitter output.

84. R-f Load Impedance. A transmission line usually is used to couple the transmitter to the antenna or to the antenna-matching unit. For this reason, the transmitter is required to match the specified transmission line with a certain range of standing-wave ratio permitted. An example is a 50-ohm line with up to 2:1 SWR. This means that this transmitter must be capable of matching any load impedance within the 2:1 SWR circle shown in Fig. 100.

85. R-f Envelope Observation. An oscilloscope set up for displaying the envelope of the transmitter output is a very useful test instrument. The r-f signal is coupled directly to the vertical-deflection plates by means of a tuned circuit and a coupling link to the transmitter. The linear sweep in the oscilloscope is used for observing most test signals. The keyed wave shape of a c-w transmitter can be observed directly. Its transient characteristics can be measured directly on the face of the oscilloscope. Also, the make-and-break characteristics which determine key clicks can be observed.

The modulation characteristics of an a-m transmitter can be observed in the same manner. The nature of appreciable amounts of distortion is readily apparent.

The envelope of an SSB transmitter with a two-frequency test signal will show the nature of large amounts of intermodulation distortion. Instead of using a linear sweep, a better observation can be made using an envelope detector for the horizontal signal. A trapezoidal figure results, and straight sides indicate low distortion. The envelope detector feeding the horizontal oscilloscope amplifier must receive its r-f input ahead of the principal source of distortion. A linearity tracer which uses two envelope detectors produces a plot of r-f input vs. r-f output of the amplifier under test. A straight line indicates low distortion. These observations are helpful in establishing proper operating condition and during "trouble shooting" but are not suitable for making measurements.

Fig. 100. Impedance locus of 2 to 1 SWR.

86. Primary Power Requirement. The nominal power-source voltage and power consumption are always given in transmitter specifications. A range of voltages may be given also, but unless the transmitter has an internal primary-voltage regulator, it should be operated as near the nominal design voltage as possible. Tube life is reduced greatly by high filament voltage. R-f power output may fall below its rated output at low line voltages. Best performance, along with long life and better reliability, will be obtained by providing a supply as near as possible to the design center value.

Alternating-current power requirements are expressed as watts with a stated power factor, or as volt-amperes. The power consumption for several operating conditions, including stand-by, may be stated. The installation wiring must be capable of delivering the maximum va or kva required by the equipment.

CHAPTER 19

RADIO RECEIVERS

By Charles E. Dean[1]

1. Classification. Radio receivers may be classified[2] according to their operating principles:

1. Tuned radio frequency.
2. Regenerative.
3. Superheterodyne.
4. Superregenerative.

Receivers of all these types have been made and used for amplitude-modulated (a-m) signals. For frequency modulation (f.m.) the superheterodyne is almost always used.

The following sections give general descriptions of these types, after which the various types and related topics will be discussed in detail.

Table 1. Designations of Frequency Bands by Abbreviations, Numbers, and Metric Classification of Wavelengths

Abbreviation	Meaning	Frequency range	Band No.	Metric designation by wavelength*
v-l-f	Very-low-frequency	3–30 kc	4	Myriametric waves
l-f	Low-frequency	30–300 kc	5	Kilometric waves
m-f	Medium-frequency	300–3,000 kc	6	Hectometric waves
h-f	High-frequency	3–30 Mc	7	Decametric waves
v-h-f	Very-high-frequency	30–300 Mc	8	Metric waves
u-h-f	Ultra-high-frequency	300–3,000 Mc	9	Decimetric waves
s-h-f	Super-high-frequency	3,000–30,000 Mc	10	Centimetric waves
e-h-f	Extremely-high-frequency	30,000–300,000 Mc	11	Millimetric waves

Other frequency band designations commonly employed are as follows:†

K band	30 kMc or 1 cm
X band	10 kMc or 3 cm
S band	3 kMc or 10 cm
L band	1 kMc or 30 cm
P band	1/3 kMc or 10 m

* In each band the wavelengths extend from 1 to 10 of the stated unit; *e.g.*, band 5 includes wavelengths from 1 to 10 km.

† The order in which these designations occur may be remembered by "*K*ing *X*erxes *S*educes *L*ovely *P*rincess."

Receivers for television service have various special features and are described in Chap. 22.

In the discussion of receivers in the present chapter, reference is often made to the various frequency bands by the abbreviations in Table 1.

[1] Hazeltine Research Corporation, Little Neck, L.I., N.Y.

[2] For an elementary treatment see Jordan, Nelson, Osterbrock, Pumphrey, and Smeby, "Fundamentals of Radio," edited by W. L. Everitt, Prentice-Hall, 1942; Keith Henney, "Principles of Radio," 6th ed., Wiley, 1952; David Grimes, "Meet the Electron," Pitman, 1944; M. G. Suffern, "Basic Electric Principles," 2d ed., McGraw-Hill, 1956. Certain general reference books containing material on various aspects of radio receivers, written for the design engineer, are listed at the end of this chapter.

2. Tuned-radio-frequency Receivers. Tuned-radio-frequency (t-r-f) receivers obtain their selectivity and r-f amplification solely through the use of selective circuits and tubes which function at the frequency of the incoming signal. This type has been largely replaced by the superheterodyne.

3. Regenerative Receivers. In a regenerative receiver, the received voltage is impressed on the grid of a vacuum tube; a portion of the resultant amplified voltage is fed back to the grid in the proper phase relation to increase the applied grid voltage. This action reduces the effective resistance of the resonant circuit where the signal is applied and thereby provides considerable amplification of the received signal.

The regenerative receiver is subject to the limitations that: (1) a critical adjustment of the feedback, just under oscillation, is necessary for substantial increase of sensitivity; (2) in adjusting the equipment more or less oscillation is likely, producing disturbance to other nearby receivers tuned to the same frequency; (3) the selectivity is sharp near the carrier and so may impair the audio quality; and (4) the employment of only a single tuned circuit results in inadequate rejection of interfering signals moderately removed in frequency. This type of receiver has therefore largely passed out of use.

4. Superheterodyne Receivers. In the superheterodyne the signal voltage is combined with a local oscillator voltage and converted into a new signal with the same modulation but a lower ("intermediate") carrier frequency, which is then amplified and detected to reproduce the original modulation.

The superheterodyne has the essential components of a t-r-f receiver, but in addition it has a frequency converter and an i-f amplifier. The frequency converter consists of an oscillator and a modulator. The oscillator and the t-r-f circuits are usually tuned simultaneously by means of a gang capacitor or a gang inductor. Constant frequency difference is maintained between the oscillator and r-f circuits, through the use of either (1) a combination of fixed shunt and series capacitors in the oscillator circuit in conjunction with a gang capacitor in which all the sections are identical, or (2) a gang capacitor in which specially shaped plates are used in the various sections, or (3) suitable arrangements of variable-inductance tuning.

The i-f amplifier in broadcast receivers generally uses two or three transformers, each containing two coupled tuned circuits. With this substantial number of circuits and the fixed and favorably chosen i-f value, the i-f amplifier provides the major portion of the amplification and selectivity of the receiver. Since the characteristics of this amplifier are independent of the frequency to which the receiver is tuned, the sensitivity and selectivity of a superheterodyne are usually fairly uniform throughout its tuning range. The r-f circuits are used primarily for eliminating certain forms of interference (especially image-frequency response) to which this type of receiver is subject.

The advantages of the superheterodyne make it the most suitable type for the great majority of radio services. It may therefore be considered the standard variety. For this reason numerous aspects of it are treated in the present chapter.

5. Superregenerative Receivers. A superregenerative receiver is a regenerative receiver in which sustained oscillations are prevented by the periodic variation of the effective resistance of the resonant circuit to which the received signal is applied.

In the superregenerator, oscillations are permitted to build up repeatedly in a resonant circuit tuned to the received signal. Sustained oscillations are prevented by the application of a quenching-frequency potential to the grid of the superregenerative tube. This potential periodically affects the tube characteristics in such a way as to stop the oscillations. The quenching frequency may be supplied either by a separate oscillator or by the superregenerative tube itself. If this frequency is in or near the upper part of the a-f range, the audio system of this type of receiver is usually provided with an a-f filter to remove the quench frequency from the audio output. An r-f stage ahead of the detector will prevent energy from being transferred from the superregenerative circuit to the antenna. A signal input of 50 to 100 μv will give an intelligible

signal although an input of 500 to 1,000 μv is generally necessary to reduce the noise to a satisfactory value. Harmonics of the quench frequency beating with the received signal make interference if the ratio between signal and quench frequencies is not 100:1 or more.

The superregenerator was used extensively to receive pulse-type signals in the v-h-f band during the Second World War and is of interest as a special type of receiver. It is discussed below, starting with Sec. 159.

6. F-m Receivers. During recent years, f.m. has come into extensive use in broadcast, police, military, and other services. The carrier frequency can be varied, if desired, over a wider range than the extent of the signal components. For example, in broadcast service in the United States a set of standards has been established by the FCC, under which audio frequencies to 15 kc are handled, but the maximum swing of frequency is ± 75 kc.

The usual broadcast f-m receiver is a superheterodyne differing in the following respects from m-f and h-f a-m broadcast receivers: (1) the carrier frequencies are much higher; (2) the band width is wider to accommodate the wide swing of the signal and the drift of the local oscillator; (3) the i.f. is higher, the present standard value being 10.7 Mc; (4) the i-f amplifier may contain one or more limiter stages; (5) the a-m detector is replaced by a frequency-sensitive device to reproduce the original audio wave; (6) the audio circuit contains a deemphasis circuit to remove the preemphasis inserted at the transmitter, this feature being another antinoise measure; and (7) the audio amplifier and speaker should be (but often are not) high-fidelity types to realize the high quality of which the system is capable.

METHODS OF TESTING

7. Characteristics Rated. Receivers are rated on the basis of numerous aspects of performance, including especially the following characteristics:

1. *Sensitivity* is that characteristic which determines the minimum signal strength to which the receiver will respond. It is measured quantitatively in terms of the input voltage or power required to give a standard output. Therefore, the receiver with the smallest value of this input is the most sensitive.

In the m-f and h-f bands the necessary input is usually stated in microvolts or in the corresponding number of decibels below 1 volt. For example, a set might have a sensitivity of 10 μv or 100 db.

The statement of sensitivity in terms of power is on the basis of *available power*, which is $E^2/4R$, where E is the open-circuit voltage of the signal generator in the sensitivity measurement and R is the resistance of the dummy antenna and the signal generator which together act as the source of signal supplying the receiver. The expression of sensitivity in terms of power is the usual practice in the higher-frequency bands.

2. *Noise* produced by a receiver, after defects such as hum, poor contacts, and faulty parts are corrected, comes from thermal noise in the antenna resistance and from noise originating in the tubes. Such noise is not large enough to be important in broadcast-band receivers having medium sensitivity and provided with an r-f stage, because high voltage gain in the first coupling network and first tube increases the signal to a value great enough to make all sources of noise small in comparison. Noise may be an important factor in the case of sets not having an r-f stage, and also in the case of all receivers requiring high sensitivity, such as automobile and farm sets.

In the v-h-f and higher ranges of frequency the input resistance of tubes is so low that the early circuits of the receiver cannot raise the signal voltage high enough to make the noise negligible. The amount of resistance noise and tube noise originating within the set is therefore important, and a quantity called the "noise factor" is used as a measure of this quantity. This is described in Sec. 19.

3. *Selectivity* is the degree to which the receiver is capable of differentiating between the desired signal and signals of other carrier frequencies. This characteristic normally requires one or more graphs for its expression.

4. *Fidelity* is the degree to which the receiver accurately reproduces at its output terminals the modulation possessed by the received wave.

5. *Maximum undistorted output* for a broadcast receiver is defined as the maximum power output which can be obtained from it when the output voltage does not contain more than 10 per cent of total harmonics.

Standard methods of testing receivers have been established by the IRE, covering broadcast receivers and to some extent receivers for other services. Various ones of these standards have been approved by the American Standards Association and therefore have the status of American Standards. The following is a current list of IRE standards[1] relating to receivers:

"Standards on Radio Receivers: Methods of Testing Amplitude-modulation Broadcast Receivers: 1948," 48 IRE 17S1 and ASA C16.19-1951.

"Standards on Radio Receivers: Definitions of Terms: 1952," 52 IRE 17S1, *Proc. IRE*, **40,** 1681, December, 1952.

"IRE Standards on Radio Receivers: Method of Testing Receivers Employing Ferrite Core Loop Antennas: 1955," 55 IRE 17S1, *Proc. IRE*, **43,** 1086, September, 1955.

"Standards on Radio Receivers: Methods of Testing Frequency-modulation Broadcast Receivers: 1947," 47 IRE 17S1 and ASA C16.12-1949.

Supplement to the above, "Tests for Effects of Mistuning and for Downward Modulation: 1949," 49 IRE 17S1 and ASA C16.12a-1951, *Proc. IRE*, **37,** 1376, December, 1949.

"Standards on Radio Receivers: Open Field Method of Measurement of Spurious Radiation from Frequency Modulation and Television Receivers: 1951," 51 IRE 17S1, *Proc. IRE*, **39,** 803, July, 1951.

"IRE Standards on Receivers: Methods of Measurement of Interference Output of Television Receivers in the Range of 300 to 10,000 Kc: 1954," 54 IRE 17S1 and ASA C 16.25-1955, *Proc. IRE*, **42,** 1363, September, 1954.

"IRE Standards on Methods of Measurement of the Conducted Interference Output of Broadcast and Television Receivers in the Range of 300 Kc to 25 Mc: 1956," 56 IRE 27S1, *Proc. IRE*, **44,** 1040, August, 1956.

"Standards on Railroad and Vehicular Communications: Methods of Testing: 1949," 49 IRE 16S1, *Proc. IRE*, **37,** 1372, December, 1949.

"Standards on Electron Devices: Methods of Measuring Noise: 1953," 53 IRE 7S1 and ASA C 60.13-1954, *Proc. IRE*, **41,** 890, July, 1953.

"American Recommended Practice for Loudspeaker Testing: 1942," IRE 6S1 and ASA C 16.4-1942.

"Standards on Audio Techniques: Definitions of Terms: 1954," 54 IRE 3S1, published in *Proc. IRE*, **42,** 1109, July, 1954.

"Standards on Television: Methods of Testing Television Receivers: 1948," 48 IRE 22S1 and ASA C 16.13-1949.

The following paragraphs give some of the essentials of the methods specified in these standards.

8. Definition of Terms.

1. *Sensitivity, selectivity, fidelity,* and *maximum undistorted output* were defined in Sec. 7.

2. *Normal Test Output.* An a-f power output of 0.5 watt in a standard dummy load connected across the output terminals of the receiver is the normal test output of a home receiver. This value applies unless the limited maximum undistorted output of the receiver qualifies it for one of the smaller values of normal test output given in Table 2. For automobile receivers, however, the value is 1.0 watt.

Table 2. Normal Test Output for Various Classes of Receivers

Max Undistorted Output, Watts	Normal Test Output, Watts
1 or more	0.5
0.1 or more but less than 1	0.05
Less than 0.1	0.005
All automobile receivers	1.0

[1] Also available is Publication 69 of the International Electrotechnical Commission, "Recommended Methods of Measurement on Receivers for Amplitude-modulation Broadcast Transmissions," 1954; sold at $3 by American Standards Association, 70 East 45th St., New York 17, N.Y.

Normal test output is also called standard test output.

3. *Sensitivity-test Input.* This is the rms r-f voltage, modulated 30 per cent at 400 cycles, which results in normal test output at resonance. It is applied through a standard dummy antenna.

4. *Selectance* is the selectivity against interference on a definite channel near the one to which the receiver is tuned. As shown in Fig. 1, it is expressed in decibels, or it may be given as the corresponding voltage ratio. The selectance against the next channel (10 kc from resonance) is also called the adjacent-channel attenuation (ACA), and against the next channel beyond (20 kc from resonance) is called the second-channel attenuation

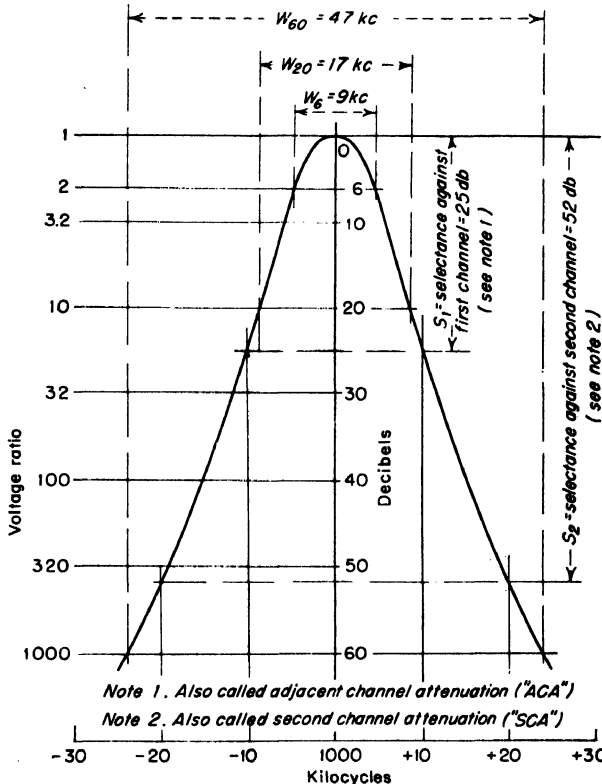



FIG. 1. Definition of selectance and band-width values on typical selectivity curve of a complete receiver.

(SCA). The values should be averages of the decibel figures for the particular frequencies above and below the resonant frequency, or the corresponding voltage ratio obtained by taking the geometric mean of the two observed voltage ratios. Notation to designate each observation has been established, S_{+1} representing the selectance for the first channel above the tuned frequency and S_{-1} the first channel below. The letter S_1 represents the mean value computed as described. Likewise S_{+2}, S_{-2}, and S_2 represent values of the SCA.

5. *Band width* is another means of expressing selectivity and is used extensively. It is the total width of the frequency band at the edges of which the attenuation has a specified value. In many laboratories the notation W_6, W_{20}, W_{40}, and W_{60} is used to represent the band width at 6, 20, 40, and 60 db, respectively.

9. Receiver Test Equipment.

1. As a source of voltages for testing, a signal generator is necessary. This is a shielded oscillator whose frequency, for testing a-m receivers, can be varied over the range of carrier,

intermediate, and image frequencies of interest. Often provided is the range from 100 to 30,000 kc. Should it be desired to make fidelity measurements, an a-f oscillator is provided to modulate the r-f oscillator by a known amount at any frequency from 30 to 10,000 cycles. For most tests 400-cycle modulation is used. A calibrated attenuator is provided to furnish a range of outputs from 0.3 μv to at least 100,000 μv and preferably to 2,000,000 μv (from 130 db below to a preferable value of 6 db above 1 volt).

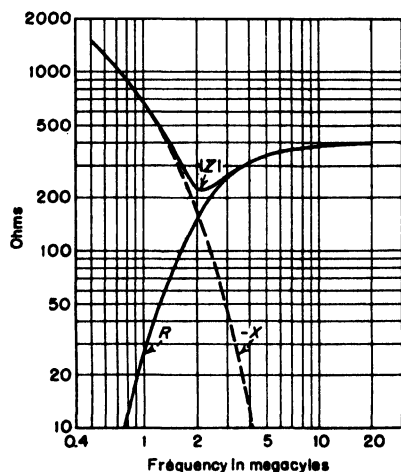
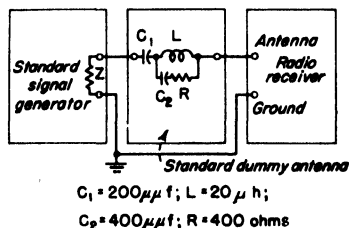



FIG. 2. Construction, use, and impedance characteristics of standard dummy antenna representing open antennas for broadcast reception in the m-f and h-f bands.

For the v-h-f and higher ranges, special signal generators are required. For the f-m broadcast band of 88 to 108 Mc, the feature of f.m. is, of course, necessary in addition to a suitable range of carrier frequencies.

2. *Standard Dummy Antenna.* The standard dummy antenna for a broadcast radio receiver is shown in Fig. 2, which is taken from the 1948 IRE receiver standards. It simulates typical actual outdoor antennas from 150 to 30,000 kc. The resistive and reactive components of the impedance are shown, as well as the modulus, which is the square root of the sum of the squares of the two components. At the higher frequencies the reactance can be seen to become negligible and the resistance and impedance magnitude to approach a value of 400 ohms, which is representative of unbalanced broadcast antennas in this frequency range.

For automobile receivers, dummy antennas of various types have been used. If the set is intended for a particular car and antenna, a dummy representing those conditions should be devised and used. One general dummy design consists of 7 ohms in series with 160 μμf, the latter shunted by 80,000 ohms. Another design consists of 100 μμf in series with 20 ohms. Another is merely a 40-μμf capacitor.

For sets with built-in loop antennas, the usual source of test signal is a fixed loop producing the test field. This is described in Sec. 10.

For the higher frequencies special cable matching the generator output circuit is desirable.[1] Also attention to the balanced or unbalanced relation to ground of generator and receiver may be needed, especially if the simultaneous use of more than one signal generator is involved.[2] Television and f-m broadcast receivers normally receive signal from a balanced 300-ohm transmission line. The dummy antenna for these cases is therefore a 300-ohm resistor, of which the signal generator will furnish a part. The resistance added in the circuit as the dummy antenna is ordinarily divided into two equal parts, one connected to each high terminal of the receiver.

3. *Standard Dummy Load.* This is a pure resistance equal to the 400-cycle impedance of the voice coil of the loud-speaker supplied with the receiver. The load resistor should be capable of dissipating the maximum power output of the receiver without an appreciable change in resistance.

4. *Harmonic-measuring Circuit.* For this purpose a harmonic analyzer capable of measuring frequencies up to 10,000 cps can be used. The instrument should have sufficient selectivity to separate each harmonic and should give the amplitude of each within an accuracy of 5 per cent. However, in normal practice, test equipment is generally used by which a composite test is made evaluating the total harmonic content as a single figure.

[1] Peterson, Arnold, Output Systems of Signal Generators, *Gen. Radio Experimenter*, June, 1946.

[2] Rankin, J. A., Receiver Input Connections for U-H-F Measurements, *RCA Rev.*, **6**, 473, April, 1942; the observations were made at 43 Mc.

10. Sensitivity Test. The sensitivity of a receiver designed for an open antenna is determined by impressing an r-f voltage of 400-cycle 30 per cent modulation in series with a standard dummy antenna and adjusting the intensity of the input voltage until normal test output is obtained at resonance for various carrier frequencies.

A receiver provided with an air-core loop antenna is tested by introducing signal inductively from a coaxial coil fed by the signal generator, or by introducing the test voltage conductively in series with the loop antenna. With the driving coil, which is the usual method, the field strength at the loop antenna is

$$E = 18.85NA^2 \frac{I}{X^3}$$

where E = equivalent electric field intensity, μv per m
N = number of turns in driving coil
A = radius of driving coil, cm
I = current in driving coil, ma
X = distance between centers of driving coil and loop antenna, m

The value of X should be at least twice the maximum dimension of either the driving coil or antenna, and much less than the distance to surrounding objects, such as the walls of the screened room.[1]

In the frequent cases of a ferrite-core loop antenna, the point to be considered the center of the receiver loop may be uncertain. For this and other reasons a modification of the air-core-loop procedure was adopted by IRE in 1955 and the standard embodying this is listed in Sec. 7. In this procedure the coaxial arrangement of the two loops is replaced by an arrangement in which the two axes are parallel; in particular, the center of the ferrite core is placed in the plane of the transmitting loop at a distance of 24 in. Other conditions are specified to remain the same, with the result that the field strength at the receiving loop in microvolts per meter can be taken as $\frac{1}{20}$ of the indicated output voltage of the signal generator; or in decibels, the field strength expressed in decibels below 1 volt per m is 26 db below the generator output expressed in decibels below 1 volt. A check to ascertain if the screen room being used is large enough to give free-space conditions, or to determine the necessary corrections of data, is recommended.

Detector sensitivity is a measure of the sensitivity of that part of the receiver including the detector and the audio amplifier. It is expressed in decibels below 1 volt, or as the equivalent fraction of 1 volt, and states the input of modulated i-f voltage necessary for normal test output. It is often desirable to obtain the result indirectly by measuring the sensitivity on the last i-f grid and correcting this value with the separately measured gain of the last i-f stage.

11. Selectivity Test. Selectivity is determined by tuning to the test frequency, observing the sensitivity, and then, with the receiver adjustments undisturbed, measuring the r-f input signal necessary to give normal test output at other frequencies. Points usually recorded are the discrimination at ±10 and ±20 kc (which are called *selectance values*) and the band width at 6, 20, and 60 db. In most cases it is not necessary to disable the a-v-c system because the levels at the output of the receiver and at the detector are constant, whence the a-v-c bias is constant.

In this determination of selectivity, only one frequency is applied to the receiver while each observation is made, and the test is therefore sometimes called the *single-signal* method of measurement of selectivity. It affords a measure of the rejection of unwanted signals by the frequency discrimination of the tuned circuits of the receiver, but it gives no information on effects which interfering signals can produce as a result of nonlinear tube characteristics or a result of the interference influencing the magnitude of the a-v-c bias. The term "effective selectivity" is sometimes used to indicate the over-all ability of a receiver to resist interference.[2]

[1] Practical procedures and designs of driving coils for testing air-core-loop receivers can be found in W. O. Swinyard, Measurement of Loop-antenna Receivers, *Proc. IRE*, July, 1941, p. 382.

[2] "Recommendation No. 95," Documents of London Assembly of International Consultative Committee on Radio ("CCIR"), I, 105–117, 1953. Sturley, K. R., "Radio Receiver Design," Part II, pp. 272–278, Wiley, 1945.

Tests of effective selectivity can be made by the use of two signal generators, one to represent the desired signal to which the receiver is tuned and the other to represent the interference. The 1948 IRE standards, listed in Sec. 7, include the single-signal selectivity test and also three two-signal tests, these being, respectively, for cross talk interference, whistle interference, and blocking interference. *Cross-talk interference* is manifested by the transfer of modulation on the undesired carrier to the desired carrier; this is caused by the nonlinearity (or overloading) of an early tube of the receiver and consequent modulator action. *Whistle interference*, consisting of "tweets," can be thought of as resulting from interaction between harmonics of the undesired carrier and of the oscillator of the receiver; they are further discussed in Secs. 50 to 52. *Blocking interference* acts by way of the a-v-c system or by overloading to cause the desired-signal output of the receiver to depend on the interference, the effect being ordinarily to reduce the output of the desired signal.

12. Electric Fidelity Test. This is determined by tuning the receiver to 1,000 kc, maintaining the input at 5,000 μv and 30 per cent modulation, varying the modulation frequency from 30 to 10,000 cps, and taking readings of relative output voltage at convenient modulation frequencies. The results at various a.f.s are expressed in terms of the value at 400 cps, which is taken as zero decibels or 100 per cent.

If the receiver has one or more tone controls, the test should be made with settings of the tone controls which will give maximum and minimum response at high and low frequencies.

13. Tests of Image Rejection and I-f Rejection. The ability of a superheterodyne to suppress the undesired image signal is measured as in the sensitivity test except that the signal generator is adjusted in frequency to deliver the image signal. This value of image sensitivity is then compared with the regular sensitivity to compute the image rejection, which is expressed in decibels or as a voltage ratio.

Any superheterodyne is more or less subject to interference from a station transmitting on or near the i.f. of the set. The receiver is tuned to the r.f. of interest and the signal generator to the i.f., *e.g.*, 455 kc. In this way the i-f sensitivity is measured. The i-f rejection is obtained by comparison with the normal sensitivity for the particular r.f.

Tests of other spurious responses are described in Sec. 52.

In the case of f-m receivers, the sensitivity to i-f interference at the antenna terminals of the set is usually measured with an f-m signal of $\pm$22.5-kc swing varying at a rate of 400 cps. The use of f.m. for the test signal determines the ability of the front-end circuits of the receiver to reject this spurious signal and avoids the test result depending on the behavior of the detector and the limiter, if any. Further treatment of tests on f-m receivers is given in Sec. 22.

14. Maximum Undistorted Output. This is determined by increasing the output in successive steps and measuring both the power output and the percentage of harmonics. The maximum undistorted output is the least power output which contains a total rms harmonic voltage content of 10 per cent of the fundamental voltage.

15. Maximum Output and A.V.C. For the maximum output the manual volume control of the receiver is set at maximum and a 1,000-kc signal with 30 per cent 400-cycle modulation is applied in increasing intensity. For open-antenna receivers the range of input signal is from 120 to zero db below 1 volt (1 μv to 1 volt). With air-core-loop receivers the observations are similar except that the intensity is in microvolts per meter and the maximum obtainable with the usual laboratory equipment is 200,000 μv per m, or 14 db below 1 volt per m. For ferrite-loop receivers the upper limit of observation is 100,000 μv per m, or 20 db below 1 volt per m. The output using either open or loop antenna is plotted as a curve. According to latest practice, additional curves are taken with zero, 10, and 80 per cent modulation.

To show the a.v.c., 30 per cent modulation is used and the manual volume control is reduced so that the power output for maximum signal input is $\frac{1}{10}$ to $\frac{1}{2}$ the largest value in the preceding tests. Again input signals over the full range of intensity are applied and a plot made of the observed output. A superior a-v-c system gives a

curve which, with increasing signal intensity, rises rapidly at low intensities and is then relatively flat over the medium and high range. The a-v-c figure of merit is defined as the reduction of the signal below a certain reference level for which the output declines only 10 db. For receivers intended for a wide range of signal input voltages (such as home receivers using open antennas), the reference level for determining the a-v-c figure of merit is 20 db (100,000 μv). For loop-antenna sets it is 26 db below 1 volt per m (or 50,000 μv per m). For receivers or tuning ranges primarily intended for small values of signal input, such as automobile receivers and the h-f bands of broadcast receivers, this reference level is 46 db (5,000 μv).

16. Hum Tests. Hum originating in a-f circuits is called "residual hum" and is measured with no r-f input to the detector. The i-f amplifier is disabled by shunting $\frac{1}{10}$ μf from the last plate to ground. Measurements are made with the volume control at its lowest and highest positions. In case there is a maximum of hum at an intermediate setting of the volume control, a measurement is made here also. The tests are normally made with an output voltmeter and may be computed and recorded in microwatts.

However, if the disconnected speaker shows apreciable hum voltage at its terminals, because of the presence of hum fields, the measurement should be made in terms of speaker current with the speaker normally connected; in this case the current-measuring equipment should not appreciably disturb normal relations in the circuit.

Hum produced by modulation of the carrier in the receiver is called "modulation hum" and is measured if it is several times as great as the residual hum. Values of modulation hum smaller than this cannot easily be measured with sufficient accuracy to be useful. For measuring modulation hum, signals at 1,000 kc of 50, 5,000, 100,000, and 1,000,000 μv and of 400-cycle 30 per cent modulation are applied, with the volume control adjusted at each signal intensity to give approximately the maximum undistorted output. At each signal intensity the observations consist of noting the audio voltage with the 30 per cent 400-cycle modulation and then switching off the modulation and noting the voltage of each hum component. The results are converted to per cent modulation by taking the ratio of hum-component voltage to 400-cycle voltage and multiplying by 30. For example, if the voltage of a particular hum component is 0.1 volt and the 400-cycle voltage is 1.0 volt, the noise modulation for this hum component is 3 per cent.

A simplified single measurement of either residual or modulation hum is often made by observing the effective total of all hum components. In this way a single power or a single hum-modulation-percentage value is obtained.

17. Tests Involving Noise. The 1948 IRE Standards include a measure of the amount of the random types of noise originating in a receiver. This is called "e.n.s.i.," standing for *equivalent noise side-band input*. The result is expressed in microvolts, or decibels below 1 volt, at the antenna terminal. Typical values for open-antenna receivers of the home and automobile types exceed 110 db. For such home receivers operating in the m-f band, with an r-f stage and typical antenna-coupling voltage gain of 15 to 18 db, the e.n.s.i. normally amounts to 121 to 123 db. Automobile receivers with antenna voltage stepups as high as 26 db have typical e.n.s.i. of 132 to 137 db. An open-antenna home receiver without an r-f stage has increased noise because converter tubes are more noisy than amplifier tubes, and such a set may be expected to have an e.n.s.i. of about 117 db.

The significance of e.n.s.i. may be appreciated when it is noted that satisfactory reproduction of a broadcast program for the average listener requires a signal-to-noise ratio of 40 db, and mediocre reproduction requires a ratio of 20 or 25 db. For example, a set with an e.n.s.i. of 117 db requires for satisfactory reproduction a signal strength of 77 db, whereas a set with an e.n.s.i. of 137 db will give equally good reproduction on a much smaller signal, *viz.*, one of 97 db below 1 volt.

Another quantity describing noise performance is the *noise factor*. This has come into wide use, especially for receivers operating in v-h-f and higher frequency ranges. A discussion of noise factor is given in Sec. 19.

18. E.N.S.I. Test. The procedure for measuring e.n.s.i. is as follows: (1) select a convenient carrier of about 100 db below 1 v (10 μv) and call this $(E_s)_{db}$; (2) with this carrier applied without modulation, measure the rms noise voltage delivered by the receiver, calling this E_n'; (3) switching on 400-cycle 30 per cent modulation and inserting a 400-cycle band-pass filter in the output circuit of the receiver (which will remove practically all the noise), measure the rms 400-cycle output, calling this E_s'; (4) compute e.n.s.i. from the formula

$$(E_n)_{db} = (E_s)_{db} + \left(\frac{E_s'}{E_n'}\right)_{db} + 10$$

Example: Consider a set measured at 100 db and giving signal and noise outputs of 50 and 10 volts, respectively. The output signal-to-noise voltage ratio is 5, or 14 db. The e.n.s.i. is then 100 + 14 + 10 = 124 db.

If E_n and E_s are in microvolts and E_s' and E_n' in volts, the e.n.s.i. in microvolts can be obtained from the relation

$$E_n = 0.3E_s\left(\frac{E_n'}{E_s'}\right)$$

Example: To illustrate the application of this formula, consider the example just given. The value of E_s is 10 μv, corresponding to 100 db, and E_n'/E_s' is $^{10}/_{50}$ or 0.2, whence $E_n = 0.3 \times 10 \times 0.2 = 0.6$ μv, which is the same as 124 db.

E.n.s.i. is an indication of the amount of thermal and tube noise referred to the antenna terminal. It is not intended to show the amount, or the presence, of other types of disturbance, such as hum, atmospherics, or noises due to faulty operation. It is not important with insensitive sets because they have insufficient amplification to build up the thermal and shot noise to appreciable magnitude. The significance of the name "equivalent noise side-band input" is this: a single side frequency 400 cycles removed from the carrier and having this (e.n.s.i.) strength would produce an rms output of 400 cycles just equal to the rms value of the noise output produced by the receiver. The normal a-m signal has an upper side frequency and a lower one with such phase relations that in the detector the audio voltage due to the beating of the upper side frequency with the carrier adds to the audio voltage due to the beating of the lower side frequency with the carrier; *i.e.*, the total audio power of E^2/R is quadrupled in comparison with the value it would have if only one side frequency were present. However, with noise the side frequencies are random in phase, with the result that the audio output from components of higher frequency than the carrier adds in terms of power to the output from components of lower frequency, a doubling rather than a quadrupling of the power. With noise there is an absence of cooperation between the upper and lower side bands, and this result is represented by considering the e.n.s.i. to be a single side frequency.

E.n.s.i. and especially noise factor, which is presented below, have become matters of importance with receivers for the v-h-f and higher frequency ranges. One reason is the limited tube input impedances, because these limit the voltage gain which can be realized in coupling networks, with the result that the early coupling networks of the receiver cannot increase the voltage of a weak signal sufficiently to make it amply override the random noise produced in the early resistances and tubes of the receiver.

19. Noise Factor. The value of the e.n.s.i. of a receiver does not indicate directly whether the noise is large or small in comparison with an ideal receiver. For this reason, another measure of noise, called the "noise factor" or "noise figure" or "excess noise increment," has come into wide use, especially for receivers designed for the v-h-f, u-h-f, and s-h-f bands. This quantity, broadly speaking, gives a comparison between the actual noise output of the receiver and the smaller noise output which would be produced if the only source of noise were from the antenna resistance. Thus, the actual receiver is compared with an ideal having no internal sources of noise and reproducing only the noise originating in the resistance of the antenna. However, from a stricter

standpoint, the noise factor is independent of the audio-amplifier characteristics and of the total gain of the receiver. It is a measure of the noisiness of the receiver resulting from the introduction of noise in the early coupling systems and tubes.

The noise factor is expressed in decibels or as a power ratio; it is not customary to state it as a voltage or current ratio. The larger the noise factor, the noisier the receiver. Typical values range from 8 to 30 db. The noise factor is of interest with any receiver having sufficient gain so that the ability to receive weak signals is limited by the presence of noise originating in the receiver.

Before giving the precise definition of noise factor, it is advisable to describe *available power*. Any source of voltage having an open-circuit value E and associated with an internal resistance R will deliver maximum power to a load of matched value, *i.e.*, a load equal to R. The load voltage is then $E/2$, and the current is $E/(2R)$. The maximum load power is the product of this voltage and current and is, therefore, $E^2/(4R)$. This is called the *available power* from the particular source.

An actual or dummy antenna of resistance R can be shown to have an open-circuit rms thermal-noise voltage equal to $\sqrt{4KTRB}$, where K is a constant with the value 1.37×10^{-23}, T is the absolute temperature (usually taken as 290°), R is the resistance in ohms, and B is the total band width in cycles between the 3-db points. The available noise power is KTB. For the noise factor, all internal noise sources are referred to the input of the receiver in terms of power, and the noise factor is the ratio which the sum of this power plus KTB bears to KTB, expressed in decibels or as a power ratio. A perfect receiver reproduces only the noise of the antenna and has a noise factor expressed as a power ratio of unity, or zero decibels.

The noise factor of receivers designed for frequencies from the m-f range up into the u-h-f range can be conveniently measured with a temperature-limited diode, *i.e.*, one without space charge, as a noise source. The noise generator and dummy antenna are connected, but the noise generator is not yet turned on. An observation is made of the noise output power P_1 of the i-f amplifier under these conditions. Then the noise generator is turned on and its output is adjusted until the noise output of the i-f amplifier has increased to $2P_1$. The direct space current I of the diode (which is the means of adjusting the noise generator output) is then recorded. It can be shown that the noise figure of the receiver in decibels is

$$10 \log_{10} (20IR)$$

where R is the resistor through which the noise current flows. For example, with 0.040 amp and 70 ohms, the value of $20IR$ is 56, corresponding to 17.5 db.[1]

For receivers operating above 200 Mc, and especially above 1,000 Mc, a suitable noise source for measuring noise factor has been found in the fluorescent lamp. A noise output of 15 or 16 db above the resistance noise is furnished by these sources. Daylight types of lamps of 6 and 8 watts are useful, and also an argon lamp designated as the EIA 6357. For precise results in the 9,000-Mc region argon lamps are recommended with the bulb temperature held between 40 and 50°C.[2]

Noise sources such as those used for measuring noise factor are in some cases very convenient for determining the sensitivity of receivers because there is no necessity of tuning the generator for each frequency. The method, however, has the obvious limitations that the receiver band width must be uniform over the range of frequencies, and spurious responses, such as the image, should be small.

There are numerous refinements to which attention should be given in making noise-factor and other types of measurements on receivers operating in the v-h-f and higher frequency ranges. The use of wave guides, with their various "plumbing"

[1] For further treatment of noise factor, see H. T. Friis, Noise Figures of Radio Receivers, *Proc. IRE*, July, 1944, p. 419. Goldberg, H., Some Notes on Noise Figures, *Proc. IRE*, **36**, 1205, October, 1948; Moxon, L. A., "Recent Advances in Radio Receivers," pp. 1–19, 73–83, Cambridge, 1949. van der Ziel, A., "Noise," pp. 32–87, Prentice-Hall, 1954.

[2] Mumford, W. W., and R. L. Schafersman, "Data on the Temperature Dependence of X-Band Fluorescent Lamp Noise Sources," *Trans. IRE*, **MTT-3** (6), 12, December, 1955.

accessories, and numerous other techniques not employed at lower frequencies gives a special character to measurements of receivers for microwave use.[1]

20. Measurement of Interference Produced by Receivers. In common with other electrical appliances, receivers can produce unwanted output which interferes with reception at other receivers. Methods of measuring interference from appliances[2] are therefore applicable to receivers, but for the most important cases special measuring equipment and procedures have been developed.

Interference produced by receivers has been most serious in the v-h-f and lower u-h-f bands and has led the FCC to set up the requirement that, after certain dates, receivers for all or any portion of the range from 30 to 890 Mc must be certificated as to compliance with certain definite requirements in this connection, as discussed below. These requirements include (1) sufficiently low radiation of interference at any frequency from 25 to 1,000 Mc, and (2) sufficiently low interfering voltage delivered to the power-supply line at any frequency from 0.450 to 25 Mc.[3]

The first one of these two requirements is of interest in the design of f-m receivers for the broadcast band of 88 to 108 Mc. The amount of radiation is measured at a distance of 100 ft with various antenna heights and procedures in accordance with the IRE standard of 1951.[4] In particular, this specification requires the antenna of the

Table 3. Maximum Radiation at 100 Ft Permitted by Federal Communications Commission for Receivers Tuning to Any Part of Band from 30 to 890 Mc

Frequency of Interference, Mc	Maximum Field Strength, μv/m
25–70	32
70–130	50
130–174	50–150 (linear interpolation)
174–260	150
260–470	150–500 (linear interpolation)
470–1,000	500

receiver under test to be at an elevation of 30 ft, taken as a practical value and a compromise between the wide extremes of actual usage. The antenna used to pick up the radiation at 100 ft is initially at a height of 20 ft. The receiver under test and its antenna, both mounted on a turntable, are then rotated and the turntable stopped at the point of largest interference reading. The pickup antenna is then gradually lowered to a height of 7 ft, and the largest field strength found is recorded as the value of the radiated interference. Separate observations are made with horizontal and vertical positions of the pickup antenna so as to obtain data for both wave polarizations. These observations are made with the receiver under test connected normally to its antenna. To test for chassis radiation, the antenna of the receiver being tested is replaced by a 300-ohm noninductive resistor and the test procedure repeated. The highest of the four values (two polarizations for antenna connected and the same without the antenna) is entered as the radiation for the particular frequency setting of the receiver under test and the particular frequency at which the field-strength meter is set. Other frequency settings of the field-strength meter are then used, and then the measurements are repeated at other tuning adjustments of the receiver under test. In terms of the frequency of the interfering radiation, the values must not exceed the amounts shown in Table 3.

[1] Wind, M., and H. Rapaport, "Handbook of Microwave Measurements," Polytechnic Institute of Brooklyn, 1954; of receiver interest are Secs. XIV through XVI, totaling 108 pages.

[2] Mortenson, K. E., and C. J. Truax, A Method of Making Screen Room Interference Measurements, *Communication and Electronics* (AIEE), No. 22, p. 746, January, 1956. Thomas, L. W., Radio Interference Measurement Techniques, *Trans. IRE*, **PGI-4,** 13, October, 1955.

[3] Federal Communications Commission, "Rules, Part 15: Incidental and Restricted Radiation Devices: February 1956 Edition," *Federal Register*, Dec. 29, 1955; also available separately from the FCC, Washington 25, D.C.

[4] "Standards on Radio Receivers: Open Field Method of Measurement of Spurious Radiation from Frequency Modulation and Television Broadcast Receivers, 1951: Standard 51 IRE 17S1," *Proc. IRE*, **39,** 803, July, 1951.

The measurement of interfering radiation requires a field-strength meter, which is a radio receiver of special design. These instruments are briefly described in Sec. 197.

21. Measures of Sensitivity Allowing for Noise. E.n.s.i. and the noise factor are measures of performance with regard to noise and do not give information as to receiver sensitivity. A need for a single quantity indicating performance with regard to both noise and sensitivity has been recognized for two decades or more in connection with military and naval receivers, and has led to the specification that the volume control of the receiver be reduced if necessary to avoid output noise exceeding a stated amount. The sensitivity of the set was then measured with this position of the volume control, and the result required to pass the sensitivity provision of the specifications. For example, in one of the armed services a receiver for headphone reception of voice was tested by adjusting the volume control to give an output of 600 μw of noise with carrier present but unmodulated. Upon application of standard 30 per cent 400-cycle modulation, the output was required to increase to 6 mw, corresponding to a signal-to-noise ratio of approximately 10 db. Some successive approximation in the test in this case is necessary until the proper volume-control setting and carrier level are found. This carrier level is then the measured sensitivity.

A slightly different procedure for obtaining a sensitivity measurement which takes noise into account has been proposed by J. M. Pettit.[1] He suggests that the quantity be called the *combined sensitivity figure*. Successive approximation is avoided by arbitrarily establishing in advance the standard noise output. With the signal generator and dummy antenna connected but the generator not operating, the volume control of the receiver is advanced until standard noise output is delivered. This setting of the volume control is thereafter not disturbed and is called the *standard gain setting*. The signal generator is then started and a regular sensitivity measurement made. This value is the combined sensitivity figure.

22. Testing F-m Broadcast Receivers.[2] The standard modulation is 30 per cent of the full 75-kc swing to either side, or ±22.5 kc. The a.f. remains at 400 cycles. The least value of carrier to give normal test output, which is the same as with a.m., is called the *maximum-sensitivity test input*, and the result of the measurement is said to determine the *maximum sensitivity*. (In television broadcast service the sound channel operates with frequency modulation, for which the full deviation is ±25 kc, so that the value for 30 per cent modulation, used in testing television receivers, is ±7.5 kc.)

Another test is made with the full 75-kc deviation. Starting with a very weak signal, the input signal level is increased and the volume control reduced, maintaining normal test output, until the output distortion, measured as the effective sum of all harmonic voltages, declines to 10 per cent. This is the *maximum undistorted output*, and the signal input at this point is said to determine the *maximum-deviation sensitivity*.

The *quieting sensitivity* is determined by the least unmodulated signal which reduces the internal receiver noise to the point where the test output rises 30 db when standard test modulation is applied to the input signal. That is, the quieting-sensitivity measurement determines the input necessary to give a 30-db signal-to-noise ratio upon the application of the standard 400-cycle 22.5-kc-deviation modulation. The receiver is adjusted for greatest sensitivity except that the volume control may be reduced to prevent audio overload. This test shows the relative freedom of the receiver from noise noticeable during pauses in the program.

These three values of sensitivity are expressed in terms of input signal. It is usually considered that the poorest one of the three determines the usefulness of the receiver.

A fourth test determines the *deviation-sensitivity input*, which indicates the amount of deviation necessary for standard test output at standard mean-signal input. The

[1] Pettit, J. M., Specification and Measurement of Receiver Sensitivity at the Higher Frequencies, *Proc. IRE*, **35**, 302, March, 1947.

[2] IRE, "Standards on Radio Receivers: Methods of Testing Frequency-modulation Broadcast Receivers: 1947," designated 47 IRE 17S1 and ASA C 16.12-1949. See also supplement on testing for effects of mistuning and downward modulation (49 IRE 17S1 and ASA C 16.12a-1951), *Proc. IRE*, **37**, 1376, December, 1949.

result of this test is expressed in kilocycles deviation, *e.g.*, 5 kc, indicating a swing of this much in each direction. The test shows whether the useful sensitivity of the receiver is limited by the audio gain. The standard mean-signal input is an available power of 90 db below 1 watt, or 1,100 μv acting through a 300-ohm dummy antenna. (For the sound channels of television receivers this test is made at a signal level of 3,200 μv.)

Observations with various dummy antennas have shown the most reliable measurements on balanced-input receivers to be obtained with a signal generator having balanced-output connections and a dummy antenna consisting of 600-ohm resistors connected to each high lead of the generator, the other ends of these resistors being joined through a 400-ohm resistor. The receiver is connected across the 400-ohm resistor, and the grounds of the signal generator and receiver are connected. Each of the 600-ohm units includes the internal resistance of one side of the signal generator, so that they must actually be somewhat less to give a 600-ohm total. If the common ground connection between the generator and receiver chassis is removed, good results are still obtained.

A simpler dummy antenna for a balanced generator consists only of two 150-ohm resistors, including the generator resistance. This gives good results if the grounds of the generator and receiver are *not* connected.

For unbalanced generators, satisfactory results are obtainable with two 150-ohm resistors, reduced as necessary to give a total of 300 ohms for the generator and the two resistors in series. It is necessary to avoid a direct connection between the grounds of the two chassis.

Two precautions have been suggested[1] in this regard: (1) the polarity of the generator output should be reversed, and (2) the power-line connection for either the generator or the receiver should be moved to a different outlet. No change in the measured sensitivity of the receiver should result from either of these two changes.

23. Typical Performance. The performance of broadcast receivers for the standard broadcast band has remained much the same while their size has been reduced by the

Table 4. Representative Characteristics of the Most Widely Sold Type of Receivers for the Standard Broadcast Band*

Characteristic	Value
Sensitivity, db below 1 volt/m	72
E.n.s.i., db below 1 volt/m	100
Band width at 6 db, kc	6.5
Same at 20 db	16
Same at 40 db	27
Image rejection, db	41
I-f rejection (455 kc), db	42
A-v-c figure of merit, db	32
Loop-antenna figure of merit, db	−5
Converter sensitivity, db below 1 volt, at 1 Mc	77
Converter gain, db	31
I-f sensitivity, db below 1 volt, at 455 kc	46
I-f gain, db	36
Detector sensitivity, db below 1 volt, at 455 kc	10
Maximum undistorted output, watts†	0.07–0.80

* Values are for operation at 1 Mc.
† Maximum undistorted output varies with type of output tube.

incorporation of miniature tubes, smaller versions of other parts, and the use of ferrite-core-loop antennas. The most widely used type of receiver has a ferrite loop and four tubes plus power rectifier, the four tubes being (1) the converter; (2) the i-f stage; (3) the detector, a-v-c, and first audio amplifier; and (4) the output audio stage. The performance of such a receiver is ordinarily about as given in Table 4. Receivers with a three-gang tuning capacitor and an r-f stage have materially better image and i-f rejections, while those with improved audio features, such as a push-pull output stage, larger output transformer, and larger speaker, afford improved fidelity and volume range of reproduction.

[1] Rankin, John A., Receiver Input Connections for U-h-f Measurements, *RCA Rev.*, April, 1942, p. 473. This gives data at 43 Mc.

The loop figure of merit, as described in Sec. 32, is stated in decibels. The fact that this is negative shows that a field strength of, for example, 1,000 μv per m produces less than 1,000 μv on the first grid.

The a-v-c figure of merit, as described in Secs. 15 and 79, is the change in input signal intensity below 50,000 μv per m necessary to produce a 10-db change in output.

The maximum undistorted power output, which is defined in Sec. 14, is the largest output which will not produce more than 10 per cent effective sum of harmonic voltages, corresponding to 1 per cent total harmonic power.

The converter sensitivity, i-f sensitivity, or the detector sensitivity is the number of decibels below 1 volt characterizing the signal that is just sufficient to produce normal test output.

NOISE IN RECEIVERS

24. Necessary Signal-to-noise Ratio. Signal and noise can be measured and compared in terms of rms voltage or current, or in terms of power, and the ratio of signal to noise expressed in decibels. For satisfactory reception of the typical broadcast program, a signal-to-noise ratio of 40 db is necessary. A value of 30 db has been established as determining the "interference test output" in certain standard receiver-test procedures. In television 30 db is about the minimum signal-to-noise ratio for a satisfactory picture. In radiotelephone service, where intelligibility is the chief criterion, 18 db is sufficient. Other service, such as radio direction finding, where the indications may be read on meters, will operate with much lower signal-to-noise ratios and may even function with less signal than noise corresponding to a signal-to-noise ratio expressed by a negative number of decibels.

Any one of numerous conditions can produce disturbance to radio reception. The cause may be external to the receiver, such as atmospherics, cosmic static, interference from unfiltered electrical appliances like elevators and shavers, ignition noise from automobiles, or interference from diathermy equipment. Internal noise can arise in the form of hum, hiss originating in early circuits and tubes of the set, and clicks or other disturbance from faulty connections, tubes, fixed capacitors, or other defective parts.

25. Sources of Noise External to Receiver. Every frequency used for radio operation is subject to interference from one or more of the following sources: (1) cosmic radiation coming from the heavens; (2) terrestrial "atmospherics" from near or distant lightning storms; and (3) man-made interference from electrical appliances of many kinds. At each receiver therefore a certain interfering level exists, and the strength of the desired signal must be sufficient in comparison to serve the required purpose. If this received signal-to-noise ratio is ample, the task of the receiver is easy; but if a marginal signal-to-noise ratio is expected or encountered, special features of system and receiver design should be considered for adoption. Such features are listed in Sec. 29 and discussed in other sections concerned with particular topics of receiver technology.

Many measurements by the National Bureau of Standards[1] can be summarized as follows: (1) By choosing suitable nonmetropolitan receiving sites, man-made interference can normally be kept below terrestrial and cosmic noise; (2) at such locations for frequencies below 15 Mc thunderstorms usually represent the principal source of noise; and (3) for frequencies between 15 and 150 Mc at such locations the noise is chiefly from cosmic sources, *i.e.*, of galactic and solar origin.

The magnitude of these external noise disturbances, as manifested in voltages picked up by an antenna, is substantial, especially at the lowest radio frequencies. With a short vertical antenna at 10 kc, plots such as Fig. 3 showed the voltage from atmospherics to be 120 to over 160 db above the resistance or thermal voltage of the antenna; *i.e.*, the antenna had an effective noise factor of this many decibels from the

[1] Crichlow, W. Q., D. F. Smith, R. N. Morton, and W. R. Corliss, Worldwide Radio Noise Levels Expected in the Frequency Band 10 Kilocycles to 100 Megacycles, *Nat. Bur. Standards Circ.* 557, 1955; for sale by Superintendent of Documents at 30 cents.

atmospherics. At higher frequencies the curves decline in more or less straight lines to the region of a few megacycles. Other plots similar to Fig. 3 show considerable variation for other times of the day and of the year, the values for 2 Mc lying between zero and 84 db with an average of 48 db for the curves for medium (No. 3) noise grade. Between 10 and 30 Mc the curves fall off steeply and pass under the line for galactic noise.

The lines for man-made and for galactic noise are seen to constitute almost a single straight line from 82 db at 100 kc to 7 db at 100 Mc. These noise sources vary little with time of the day and year.

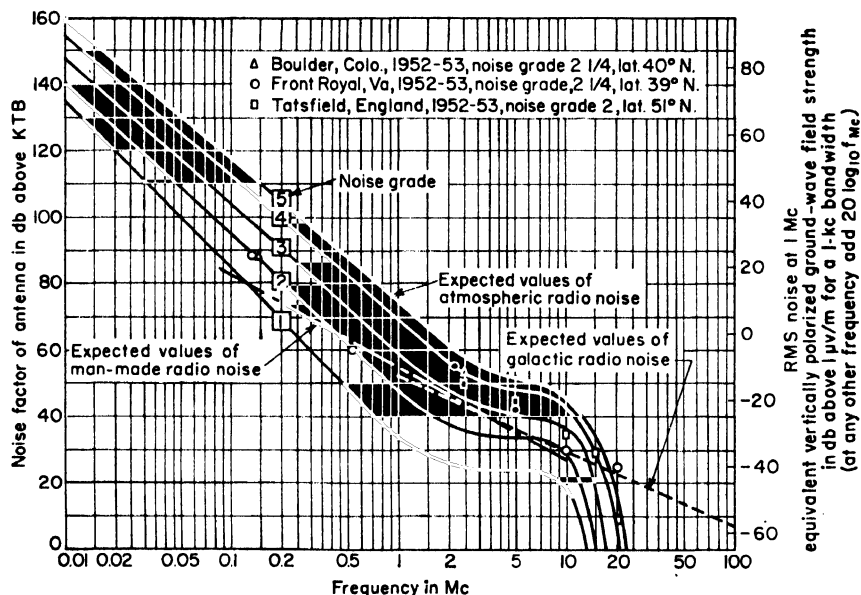



FIG. 3. Median values of atmospheric, galactic, and man-made noise expected for the hours 4 to 8 A.M. in winter, using a vertical antenna. (*From Crichlow, Smith, Morton, and Corliss, Nat. Bur. Standards Circ.* 557, 1955.)

Sources of man-made noise are very numerous, including sparking motor commutators, transmitters radiating harmonics of the assigned frequency or subharmonics coming through from the premultiplication stages, improperly controlled or shielded h-f equipment in industrial or medical service, etc.[1] Legal control of such interference has long been established in Canada[2] and more recently in the United States.

The usual methods of curing interference due to appliances[3] are (1) to connect a low-pass filter in the supply line as close as practicable to the appliance, so as to prevent the line from acting as a conductor or antenna for the interference, (2) to connect capacitors at sparking elements of appliances such as motor brushes, and (3) to locate the antenna where the noise is not encountered and use a leadin having a balanced arrangement or shielding, or both. Assembled power-line filters for various applications are offered by firms manufacturing fixed capacitors.

Automobiles and trucks will produce interference over a wide frequency range

[1] Sterling, G. E., Blights on the Radio Spectrum, *Trans. IRE*, **PGBTR-6**, 42, April, 1954.

[2] Browne, G. C. W., Control of Radio Interference in Canada, *IRE Convention Record*, **3**, Part 7, 130, 1955.

[3] "Radio Interference Suppression Techniques," U.S. Camp Coles Signal Laboratory, 1953, 270 pp.; sold as PB-111,611 by U.S. Department of Commerce, Washington 25, D.C., at $6.75. Smith, W. C., Radio Interference Suppression, *IRE Trans.*, **CS-3** (1), 8, March, 1955.

if not provided with suppression devices such as are often installed to permit radio reception on the vehicle.

Ignition interference is an example of *impulse noise*, which is to be distinguished from fluctuation noise. Impulse noise consists of discrete separate pulses with relatively long quiet intervals between. Each pulse shock-excites the usual radio tuning circuits. The resulting oscillations have the frequency of the various tuned circuits and die out exponentially; *i.e.*, each disturbing pulse produces a damped train of oscillations. Ordinarily, this train dies away considerably before the occurrence of the next pulse. Fluctuation noise, on the other hand, is always present with an instantaneous amplitude determined by statistical probability laws. Examples of it are resistance noise and tube noise, mentioned below.

26. Noise Sources in the Receiver. After noise due to poor contacts, defective parts, and microphonics has been removed from a receiver, it will, with the usual a-c power supply, still deliver noise in the form of *hum.*[1] By good design and maintenance work, this can be reduced until it is no longer appreciable. Whenever hum is encountered, special attention should be paid to the low-level audio stages (where the trouble may be heater-cathode leakage), the possibility of inadequate plate-supply filtering, or magnetic induction from stray fields. Another type of this disturbance is "modulation hum," which is present only while a carrier is being received. This is produced by nonlinearity in the r-f and i-f stages plus a hum voltage, so that the hum modulates the carrier. Such hum would, of course, be reproduced by a perfect detector and audio system and must be corrected by removing the cause in the r-f or i-f amplifier.

Other noise sources in the receiver include those of the random-fluctuation type producing a characteristic hisslike sound in reproducers and "grass" on oscilloscopes. These sources include resistance noise and tube noise, important in high-gain receivers and in receivers for the v-h-f and higher frequency ranges.

27. Fluctuation Noise. The irregular motion of electrons in any resistor and in the space current of any tube gives rise to the fluctuation type of noise.[2] These random motions result in small fluctuations of voltage and current. The instantaneous value of this voltage or current obeys the probability law, under which there is a certain probability that the value will fall in any range of interest. For example, there is a very small probability that a very large value will occur, so that the usual concept of peak amplitude does not apply. Instead, the probability of a large amplitude increases as the band width of the equipment is increased. There is no definite wave form. However, there is a definite rms value and also a definite average value for all positive or all negative amplitudes. These two averages are equal and each amounts to 80 per cent of the rms value.

The instantaneous amplitude in one direction or the other exceeds the rms value 32 per cent of the time, but it exceeds four times the rms value for only 6.4/100,000 of the time.

Another characteristic of fluctuation noise, before it has gone through any selective circuits, is that it has a uniform frequency distribution; *i.e.*, a 1,000-cycle band selected in the audio region has as much rms voltage as a 1,000-cycle band at 1, 10, or 100 Mc.

Fluctuation noise establishes a limit to the amount of amplification that can be usefully employed. At the input to a receiver or amplifier, the signal must be large enough so that it will not be lost in the noise associated with resistances and tubes in the low-power portion of the equipment.

28. Thermal Noise and Tube Noise. The random motion of electrons, under the molecular agitation due to temperature, produces noise across the terminals of any

[1] Langford-Smith, F., "Radiotron Designer's Handbook," 4th ed., published in Australia, 1952; sold in United States by Radio Corporation of America, Harrison, N.J.; on hum see pp. 1196–1201 and 1239–1241.

[2] Landon, V. D., The Distribution of Amplitude with Time in Fluctuation Noise, *Proc. IRE*, February, 1941, p. 50. van der Zeil, A., "Noise," Prentice-Hall, 1954. Pierce, J. R., Physical Sources of Noise, *Proc. IRE*, **44**, 601, May, 1956. Bennett, W. R., Methods of Solving Noise Problems, *Proc. IRE*, **44**, 609, May, 1956, gives a substantial introduction to the theory of the subject. Bennett, W. R., series of five articles on various phases of noise, *Electronics*, **29**, as follows: 154, March, 1956; 134, April, 1956; 162, May, 1956; 154, June, 1956; 148, July, 1956.

resistor. A parallel resonant circuit or other passive network manifests the same effect to the extent corresponding to the resistive component of the impedance. The magnitude of the open-circuit rms voltage is given by the relation

$$E = \sqrt{4KTRB}$$

where K = Boltzmann constant = 1.37×10^{-23}
T = absolute temperature, °K (about 300 for room temperature)
R = resistance, ohms
B = band width, cps

For radio purposes it is more convenient to express this in microvolts, ohms, and megacycles. Making these changes and inserting the numerical values of K and T, we obtain

$$E = 0.13 \sqrt{RB}$$

Values of E for wide ranges of resistance and band width are plotted in Fig. 4.

For the usual noise computations, which have a limited accuracy, the determination of B will be easy. For a selectivity curve having a shape like that of a single tuned circuit, the value of B is the total frequency interval between the two points where the curve is 3 db down. For more involved cases, a graphical determination can be made by using graph paper with linear graduations in both directions and plotting frequency as abscissas and power (or voltage squared or current squared) as ordinates. Measure the area under the curve and determine the width of a rectangle of the same area having a height equal to the highest single ordinate. The width of this equivalent rectangle is the desired value of B.

An important quantity associated with the thermal noise voltage is the *available power* from such a source. As with any source of open-circuit voltage E acting through an internal resistance R, the maximum power which can be delivered, defined as the available power, is obtained by matching the load resistor to the internal resistance. The value of the available power is $E^2/4R$. Applying this to the case in hand, the available thermal-noise power is $(0.13)^2RB/4R$, or $W = 0.004B$. Here B is in megacycles, and W is in micromicrowatts. Table 5 gives values computed from this formula. An alternative expression for available noise power is 4×10^{-21} watt per cycle of band width.

Fluctuation noise originates in tubes due chiefly to (1) the random nature of emission at the cathode and (2) the irregular and changing distribution of current between plate and screen if the tube has a screen. The first of these was discussed by Schottky[1] in 1918, along with resistance noise, all under the name *schroteffekt* (German for "shot effect") from the similarity of the sound to that of many small pellets falling on a hard surface. During the years this term has come to be restricted by most workers to tube noise resulting from random cathode emission so that resistance noise is not included. The name "partition noise" is often applied to tube noise resulting from the irregularities in distribution of current between plate and screen.

Table 5. Available Power of Thermal Noise for Various Band Widths

Band width, B, Mc	Available power, W, $\mu\mu$w	Decibels below 1 watt
10^{-4} (100 cycles)	4×10^{-7}	184
10^{-3} (1,000 cycles)	4×10^{-6}	174
10^{-2} (10 kc)	4×10^{-5}	164
10^{-1} (100 kc)	4×10^{-4}	154
1	4×10^{-3}	144
10	0.04	134

A common method of stating tube noise is to specify an equivalent resistance which, if connected at the grid and if the tube itself were noiseless, would produce the same noise output due to the room-temperature thermal noise in the resistance. This is

[1] Schottky, W., Spontaneous Current Fluctuations in Various Conductors, *Ann. Physik*, **57**, 541, Dec. 20, 1918.

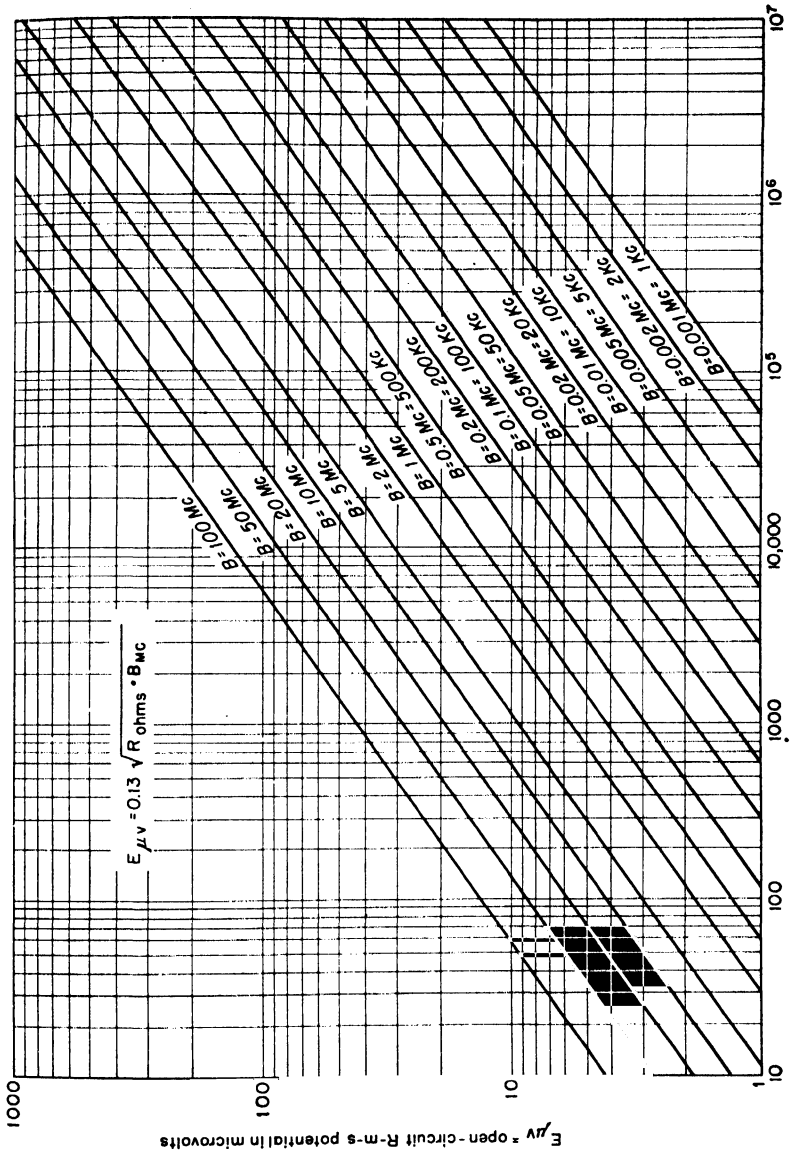



Fig. 4. Thermal noise for various resistances and band widths.

more convenient than a voltage statement because the resistance value does not depend on band width while the voltage value does.

The noisiest tube condition is with temperature saturation, *i.e.*, no space charge. In this case all electrons emitted are immediately drawn to the plate. A diode operated in this way is, in fact, the best standard source of noise for use in measuring the noise figure of a receiver. The presence of the space charge, which makes amplification possible, greatly reduces the tube noise, which is a very fortunate effect.

Values of equivalent grid resistance expressing the noisiness of various tubes have magnitudes[1] generally lying in the ranges given in Table 6.

Mixer noise increases beyond these values if the oscillator injection is insufficient. For amplifiers the effective input noise increases, but the output noise decreases as the gain is reduced by the application of more negative grid bias.

Table 6. Equivalent Grid Resistance Representing Tube Noise for Various Tube Types and Circuits

Tube and circuit	Ohms	Tube and circuit	Ohms
Triode amplifiers	200-2,300	Gradual-cutoff pentode amplifiers	2,400- 14,000
Triode mixers	900-6,000	Pentode mixers	2,800- 35,000
Sharp-cutoff pentode amplifiers	700-7,000	Hexode and heptode mixers	190,000-300,000

Tube noise is an important matter in the early tubes of receivers for the v-h-f and higher-frequency bands. Developmental triodes have been constructed with noise factors as low as 7.8 db at 500 Mc, and 9.7 db at 870 Mc.[2] For tubes operating at much higher frequencies, theory is deficient for the region of the potential minimum and from here to the cathode, but there have been useful advances in the understanding of traveling-wave tubes and the construction of low-noise models.[3]

29. Means for Obtaining Reception despite Noise. There are numerous methods for improving the operation of systems and receivers exposed to electrical noise.[4] Among these are the following:

1. Use as much antenna directivity as possible.
2. Endeavor to have a large enough antenna to intercept sufficient signal to predominate over noise sources in the receiver.
3. Reduce the band width of the receiver as much as the signal characteristics permit.
4. Have as high voltage gain as possible in the coupling between the antenna and the first tube. For this purpose select as the first tube one with as high input resistance as possible.
5. Use an r-f stage to avoid the large noise of mixers early in the circuit.
6. Consider the use of triodes rather than pentodes. In the v-h-f and higher bands consider the grounded-grid triode. At u-h-f and higher bands consider crystal rather than tube mixers.
7. Select particular types of tubes for low noise, especially for the first one or two sockets of the receiver.
8. Use preemphasis in transmission and deemphasis in reception of low-energy portions of band of modulation frequencies.
9. Use volume compression in transmission and volume expansion in reception.
10. Use f.m. or pulse modulation rather than a.m.
11. Consider the use of a squelch circuit. If impulse noise is encountered, consider noise limiters. Squelch- and noise-limiting circuits are discussed in Secs. 93 and 94.

30. External Cross Modulation. A form of interference to which all types of receivers are subject can exist where two or more stations lay down very strong signals in a given area. The trouble arises from poor contacts between various extended conductors in the vicinity, such contacts acting as frequency converters and thereby producing new r-f signals which reach nearby receivers either conductively or by radiation.

[1] See Harris, W. A., Fluctuations in Vacuum Tube Amplifiers and Input Systems, *RCA Rev.*, April, 1941, p. 505, which includes applications of extensive theoretical and practical work done with B. J. Thompson and D. O. North.

[2] Harris, W. A., Measurement and Analysis of Triode Noise, *Trans. IRE*, **ED-1** (4), 206, December, 1954.

[3] Pierce, J. R., General Sources of Noise in Vacuum Tubes, *Trans. IRE*, **ED-1** (4), 135, December, 1954. Pierce, J. R., "Electrons, Waves and Messages," Hanover House, 1956; on traveling-wave tubes see pp. 158-165 and 176-180.

[4] Albin, A. L., and H. M. Sachs, Design of Electronic Equipment for Radio-interference Reduction, *Communication and Electronics* (AIEE), No. 25, p. 352, July, 1956. Schwartz, L. S., Principles of Noise Reduction in Communication Channels, *Communication and Electronics* (AIEE), No. 23, p. 44, March, 1956. Lebenbaum, M. T., Design Factors in Low-noise Figure Input Circuits, *Proc. IRE*, **38**, 75, January, 1950.

For example, in Seattle about 1938, three stations on 920, 970, and 1,270 kc each produced strong signals in a large residential area. All three stations together were heard at 620 kc (which is 920 + 970 − 1270) as well as at 1,220 kc (920 + 1270 − 970) and at 1,320 kc (970 + 1270 − 920). At three additional frequencies, two of the three stations were heard; *e.g.*, the 920 and 970 stations were heard at 870 kc (2 × 920 − 970).

An investigation by RMA led to the suggestion that a limitation in the field strengths of overlapping coverages is advisable. A quantity called the "field product" was defined and a maximum of 0.001 recommended for it. (Values of 0.0015 to 0.005 were found in the Seattle cases where trouble existed.) Where three stations are involved, the field product is $\frac{3}{2}F_1F_2F_3$, where the Fs represent the field strengths of the three stations in volts per meter. If only two stations are involved, there are two field products, namely $\frac{3}{4}F_1^2F_2$ and $\frac{3}{4}F_1F_2^2$, and the same maximum of 0.001 applies to these.

Conductors which may form contacts and give rise to external cross modulation include house electric wiring, BX-cable sheath, antenna and ground wires, water pipes, metal lath, gas pipe, and rain spouts. The remedy is to bond the conductors together or separate them with good insulation. Whenever external cross modulation is suspected, the antenna and ground circuits of the receiver should first be checked, and then a filter tried where the power plug is inserted in the receptacle. If necessary, a new antenna with shielded leadin can be tried, and it may be helpful to install a balanced antenna input circuit.

Accounts of theory and practical observations of external cross modulation have been published.[1]

LOOP ANTENNAS IN BROADCAST RECEIVERS

The convenience of the loop antenna, in comparison with outdoor or indoor open antennas, has brought it into general use for standard-band broadcast reception. The directional properties of the loop may occasionally require the user to make a slight rotation of the entire set, since the loop is usually rigidly mounted. This has not been found objectionable.

Medium-size and large cabinets can accommodate an air-core flat loop placed an inch or more from the chassis so as to have an adequate value of Q. However, at about equal cost a ferrite rod with a compact winding can be provided, with the advantage of being suitable for placing near the chassis and in the smallest cabinets. The ferrite type is therefore used in almost all new designs of broadcast receivers for home reception.

31. Voltage Induced in Loop by Wave. The passage of a wave in the direction toward which a loop is turned induces in the loop a voltage given by the formula

$$E = 0.021\mu eNAf$$

where E = induced voltage, μv
μ = effective relative permeability (ratio of flux in coil with magnetic core to value without core)
e = field strength, μv per m
N = number of turns in loop
A = area of loop, sq m
f = frequency, Mc

The ratio E/e is by definition the series voltage in the loop divided by the field strength and is often called the "effective height" from its similarity to the effective height of an open antenna. The formula for the effective height of a loop antenna is, therefore,

$$\text{Eff. height} = \frac{E}{e} = 0.021\mu NAf$$

[1] Ebel, A. J., A Note on the Sources of Spurious Radiations in the Field of Two Strong Signals, *Proc. IRE*, February, 1942, p. 81. Eastman, A. V., and Lawrence C. F. Horle, The Generation of Spurious Signals by Nonlinearity of the Transmission Path, *Proc. IRE*, October, 1940, p. 438. Foster, Dudley E., A New Form of Interference—External Cross Modulation, *RCA Rev.*, April, 1937, p. 18.

where the effective height is in meters and the other quantities are the same as in the preceding paragraph.

The effective heights of loop antennas are small in comparison with open antennas, typical loop values for 1 Mc lying in the range from 0.003 to 0.03 m.

32. Figure of Merit of Loop. The important action performed by a loop and the coupling elements, if any, between it and the first tube is to deliver voltage to the grid of this tube. The ratio of this voltage to the field strength is usually called the "figure of merit," but is also known as the "effectiveness" and the "pickup factor." This is the best single quantity for indicating the performance of a loop and includes the effect of coupling elements, if any, such as a loading coil or transformer. In terms of Q the figure of merit is the product of Q and the effective height; therefore,

$$\text{Figure of merit} = 0.021\mu NAfQ$$

The value of the Q of an air-core loop depends greatly on whether substantial losses are introduced by close location of the loop to the metal chassis. Typical effective values of Q for air-core loops, with coupling circuits if any, lie in the range from 40 to 200.

Typical values of the figure of merit, being the product of effective height and Q, range from 0.12 to 6, which is -18 to $+16$ db. This range applies for air-core loops; ferrite-core loops are often used in quite small receivers for which performance in the lower portion of this range is satisfactory.

33. Air-core Loops. In the case of an air-core loop the designer has the option of using either a high-impedance or a low-impedance loop.

In the high-impedance design the inductance of the loop antenna constitutes all the inductive element of the first tuned circuit of the receiver. This is tuned over the broadcast band by the first section of the gang tuning capacitor. The input connections of the first tube are taken also in parallel to the loop and variable capacitor. This circuit has the advantages of simplicity and a small number of parts.

With this simple circuit the Q of the loop (reduced slightly to represent any other circuit losses) determines the selectivity of the coupling system according to the following formulas:

$$W_3 = \frac{f}{Q} \qquad W_{10} = \frac{3f}{Q}$$

$$W_6 = \frac{\sqrt{3}f}{Q} \qquad W_{20} = \frac{10f}{Q}$$

where W_6, for example, is the band width between the points at which the curve is 6 db down. Any frequency units can be used but must be the same for W and f.

In some designs it will be desired to provide the high-impedance air-core loop antenna with some means whereby an outside antenna can be connected to increase the sensitivity of the receiver for use in regions of low signal strengths.

The most common practice is to connect the antenna to one or two turns of wire around the loop, or the electrical equivalent in the form of an autotransformer obtained by tapping the loop at one or two turns from the low-potential end. However, this type of coupling has the limitation that the voltage gain varies considerably over the band. In a typical case the voltage gain was 9 db (2.8 times) at 550 kc, 21 db (11.5 times) at 1,000 kc, and 24 db (16 times) at 1,500 kc. Such gain in the middle- and upper-frequency parts of the band gives poor image rejection and may cause cross modulation, objectionable tweets, and distortion. In the particular case the coupling system had an image rejection of about 34 db at the l-f end of the band and about 27 db at the h-f end.

One simple step to improve the performance is to insert a resistance of about 500 ohms directly in the antenna connection. One such circuit had a voltage gain of 12 db (four times) at 1,000 kc.

A more thorough remedy is to use a high-inductance primary, so as to resonate the primary circuit below the broadcast band. Windings of 30 to 55 ohms d-c resistance shunted with 50 to 300 $\mu\mu f$ have been used for this purpose.

An economical and effective solution is to retain the low-inductance primary of one or two turns around the loop and add a loading coil and a damping resistor. A suitable circuit is shown in Fig. 5. The coil can be a low-cost solid-wire type. This circuit gives primary resonance either in or near the l-f end of the band, the particular point not being important because the damping by the resistor removes any difficulties in the unicontrol alignment with other tuning circuits. The performance given in Table 7 was obtained with this circuit when using the standard IRE dummy antenna.

Fig. 5. Coupling for use of outside antenna with high-impedance loop.

The i-f rejection in this design can be improved at the expense of the image rejection, if desired, by reducing the 275-μh inductance of the loading coil.

In the low-impedance design of an air-core loop, only a few turns are provided in the loop (these made as large as possible and spaced as well as possible from the chassis) and the remaining inductance of the first circuit is obtained by means of a high-Q series coil. The two may be tuned by the first section of the gang capacitor, or the receiver may be designed for permeability tuning, in which case the inductance of the high-Q series coil can be varied and the capacitance of the first circuit can be obtained from a stable fixed capacitor.

Table 7. Performance of Antenna Coupling of Fig. 5

Frequency, kc	Voltage gain, μ	Db	I-f rejection (455 kc), db	Image rejection, db
550	4.0	12.1	28	59
600	4.2	12.5	32	60
800	4.2	12.5	..	55
1,000	3.8	11.6	36	53
1,200	3.1	9.8	..	50
1,400	2.6	8.3	33	50
1,500	2.2	6.9	..	47

34. Ferrite-core Loop Antennas. The ferrites are crystals having the general chemical formula $M_n(Fe_2O_4)$, where M_n represents any divalent metal or mixture of such metals. Frequently used are nickel-zinc and manganese-zinc.[1] The magnetic properties vary from hard to soft, depending on the materials used, the completeness of crystallization, and the crystal size. In physical character the ferrites are hard and brittle, similar to ceramics, and require some care in handling to avoid breakage.

Early work had shown that powdered pressed ferromagnetic materials used as cores for loop antennas offered certain performance advantages. However, for broadcast receivers the idea was not practical until the availability of the ferrites with their higher permeability and lower loss.[2] The increased permeability is of special value because it attracts a greater flux through the loop so that a small loop area is sufficient to afford adequate signal voltage. Typical core dimensions are ⅜ in.

[1] Gorter, E. W., Some Properties of Ferrites in Connection with Their Chemistry, *Proc. IRE*, **43**, 1945, December, 1955.

[2] Grimmet, C. A., Ferrite Cored Antennae, *Convention Record of IRE*, **2**, Part 7, p. 3, 1954.

diameter and 7 in. length, and the loop has a diameter only very slightly larger with enough turns (*e.g.*, 50) to give an inductance in the region of 150 to 300 μh. The value of Q is held to below 150 or 200 to avoid critical tracking between the antenna and oscillator circuits.

The figures of merit of five ferrite loops measured recently at 1 Mc were expressed in decibels and these values averaged, giving a result of −4.9 db. Values at 0.6 Mc were 3 or 4 db poorer, and those at 1.4 Mc were higher than those for 1 Mc by 1 or 2 db. The 6-db band widths were approximately 9 kc at 0.6 Mc, 18 kc at 1 Mc, and 27 kc at 1.4 Mc.

Since a ferrite-core antenna makes very little use of electrostatic pickup, it is ordinarily much less susceptible to interference from man-made electrical disturbances than are air-core loop antennas.

Sensitivity measurements on receivers using ferrite-core loops require a different procedure from the air-core case, and an IRE standard for this purpose has been adopted.[1] Some of the details from this standard have been given above in Sec. 10.

Loop antennas are also used in direction-finding receivers, for which see Secs. 184 to 186.

35. Precautions Necessary with Ferrite-core Loop Antennas. In two respects the design of a receiver with a ferrite loop requires care which is not needed with an air-core design. One of these is to avoid the ferrite being acted on by 60-cycle magnetic fields, which would introduce modulation hum into the signal on account of the nonlinear characteristic of the ferrite. This has been encountered from the power transformer of a-c receivers and also from the clock (operating on 60 cycles) of clock radios.

The other precaution needed with ferrite loops is with respect to temperature. Above a certain temperature, which depends on the grade of ferrite being used, change of inductance will be substantial and detuning of the antenna circuit will lower the sensitivity of the receiver. The remedy is to determine the highest temperature to be encountered and to provide a grade of ferrite which will withstand this temperature.

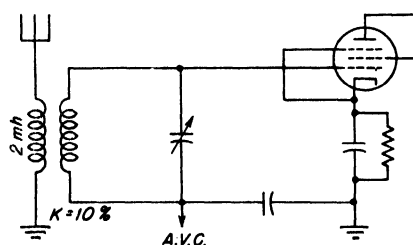



Fig. 6. Antenna transformer with high-impedance primary.

36. Antenna Couplings for Sets Accommodating Open Antenna Only. The m-f broadcast receiver designed for the weakest signals, such as in remote rural areas, must rely on a good open antenna. An effective height of 4 m, which is typical of outdoor antennas, is many times greater than obtainable with loop antennas. If few or no strong signals are expected, it is not worth while to provide a loop antenna. The design problem then is to provide good voltage gain and selectivity and also good alignment with other tuning circuits, all for any antenna capacitance within a range of 100 to 300 μμf. From a chronological standpoint this problem was solved before loop antennas became common, and the result has been in wide use for a long time. The solution is shown in Fig. 6, where the 2,000-μh primary winding is sufficient to resonate the primary circuit below 550 kc for any value of antenna capacitance. Coupling of about 10 per cent furnishes enough mutual inductance to transfer adequate signal into the secondary circuit. In the absence of any unnecessary damping, it is necessary to avoid having the primary resonant in the band, as the unicontrol alignment would be seriously affected. The relatively flat voltage-gain curve obtained with this type of transformer shows ratios of voltage on the first grid to the voltage introduced in series with the standard dummy antenna of 5.6 to 7.6 times.[2]

[1] "IRE Standards on Radio Receivers: Method of Testing Receivers Employing Ferrite Core Loop Antennas: 1955," 55 IRE 17S1, *Proc. IRE*, **43**, 1086, September, 1955.

[2] Landon, V. D., contribution to Pender and McIlwain, "Electrical Engineers' Handbook: Communication and Electronics," 4th ed., Wiley, 1950; see pp. 7-115 through 7-117.

The antenna-input system shown in Fig. 7 provides considerably greater coupling between the antenna and the first tuned circuit. This system is employed in automobile receivers where the signal intercepted by the antenna is usually quite small. By connecting a small inductance in series with the antenna so that a series-tuned circuit is formed which is resonant at approximately 2,000 kc, this system will provide a voltage gain which varies from 20 db (10 times) at 600 kc to 26 db (20 times) at 1,400 kc.

Fig. 7. Closely coupled antenna input system with common low-side capacitor.

Another antenna input system which is used extensively in automobile receivers, particularly those designed for a specific car and antenna, is to connect the antenna to a tap of approximately 30 to 50 per cent on the coil in the first tuned circuit.

Permeability tuning for the antenna and r-f interstage couplings of automobile sets is now the general rule because of the economical construction, freedom from microphonism, and convenient adaptability to push-button tuning.

TRANSISTORS AND THEIR USE IN RECEIVERS

37. Transistor Characteristics. Transistors are more costly than tubes, but have the advantages of smaller size and reduced power-supply and heat-dissipation requirements. These qualities have made transistors important devices in equipments where compact construction is essential. The small size of the transistor contributes directly to the desired miniaturization, and the reduced power requirements contribute by needing less cooling area and less space for power-supply parts.

Various types of transistors are available, and there is an extensive literature on their construction, properties, and uses.[1] In addition to the treatment in the present section, reference may be made to Chap. 10 of this handbook.

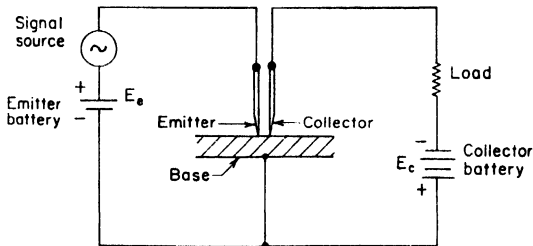



Fig. 8. Diagrammatic representation of point-contact transistor in circuit.

In common with tubes,[2] transistors have the property of producing amplification, but there are numerous points of difference relating to the mode of operation and to the proportioning of the associated circuits for effective operation.

Transistors are made of a material, particularly germanium or silicon, in the semiconductor class, whence they are often called *semiconductor devices*. From the absence of vacuum or gaseous phenomena they are also called *solid-state devices*.

The first type introduced was the point-contact transistor, having three elements as shown in Fig. 8. In this type two metallic contact points, quite close together, press against a base of semiconductor material. A signal current applied to one of the points, called the *emitter*, so affects conditions in the base that a more or less equal current flows in the circuit of the other contact point, called the *collector*.

[1] Hunter, L. P., ed., "Handbook of Semiconductor Electronics," McGraw-Hill, 1957. "Transistors I," RCA Laboratories, Princeton, N.J., 1956. Dunlap, W. C., "An Introduction to Semiconductors," Wiley, 1956. Kiver, M. S., "Transistors in Radio and Television," McGraw-Hill, 1956. Two special issues of *Proc. IRE* as follows: **40** (11), November, 1952; **43** (12), December, 1955. Coblenz, A., and H. L. Owens, "Transistors: Theory and Applications," McGraw-Hill, 1955. Herold, E. W., "Semiconductors and the Transistor," *J. Franklin Inst.*, **259,** 87, February, 1955. Shea, T. E., ed., "Principles of Transistor Circuits," Wiley, 1953. Shockley, W., "Electrons and Holes in Semiconductors," Van Nostrand, 1950.

[2] Fink, D. G., Transistors versus Vacuum Tubes, *Proc. IRE*, **44,** 479, April, 1956.

Since the internal resistance of the collector circuit is much higher than that of the emitter circuit, voltage gain and power gain are produced. It is seen in Fig. 8 that the base element forms a part of both the input and output circuits, so that this type of connection is called *common-base.* (Common-emitter and common-collector connections are also available.) There is a general correspondence of elements between transistors and tubes, with the emitter having a similarity to the cathode, the base to the grid, and the collector to the plate. In this way the common-base transistor connection corresponds to the common-grid, or grounded-grid, tube circuit; in both cases the output signal has the same polarity as the input signal.

The point-contact type of transistor was announced in 1948 and is still used, *e.g.*, in computers. However, for most purposes the later *junction types* are used, since they

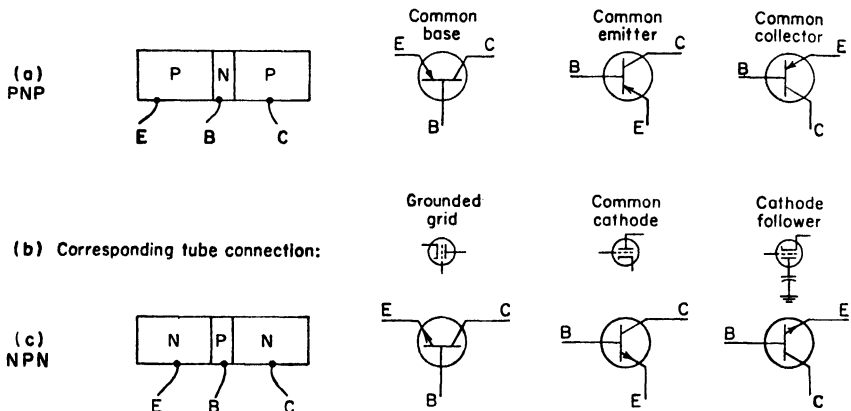



FIG. 9. *P-n-p* and *n-p-n* types of transistors with symbols for their use in the three possible classes of circuits.

have better battery economy and greater power-handling ability, and the interest of the receiver designer will normally be limited to these.

Semiconductors can be prepared to have either a surplus of electrons (in which case the semiconductor is the *n* or negative type) or a deficiency of electrons manifesting itself as "holes" which act like positive charges (in which case the semiconductor is the *p* or positive type). A junction transistor is made by creating in a single crystalline structure a region consisting of a thin layer of either type of semiconductor, acting as a base, between regions of the other type, which act, respectively, as emitter and collector. In this way *p-n-p* and *n-p-n* junction transistors are obtained. These are shown in Fig. 9. Each, like a triode tube, can be used with any one electrode common to the input and output circuits, and the usual representations for these six types of transistor circuits are included in the figure.

The operation of a junction transistor is probably most easily explained[1] in terms of an *n-p-n* type in the common-emitter connection, which comes nearest to resembling tube practice. See Fig. 10. The operation can be considered to consist of the following steps:

1. A junction between *n* and *p* types of the semiconductor operates with rectifier characteristics. This applies to the emitter-base junction and also to the base-collector junction.
2. The base battery places a forward voltage on the emitter-base junction, which causes electrons to travel from the emitter into the base.
3. The signal is added to this base bias, so that the rate at which electrons enter the base region depends on the signal amplitude.

[1] Shockley, W., Transistor Electronics: Imperfections, Unipolar and Analog Transistors, *Proc. IRE*, **40**, 1289, November, 1952. Sparks, M., The Junction Transistor, *Sci. American*, **187**, 28, July, 1952. Kircher, R. J., Properties of Junction Transistors, *Trans. IRE*, **AU-3**, 107, July–August, 1955.

4. The width W of the base is very small, of the order of 0.001 in., so that the electrons entering it mostly pass on into the collector region.

5. The base-collector junction is so biased by the collector and base batteries that electrons entering the collector region are attracted to the output terminal and pass through the load. This is the amplified output. (The base-collector junction has a backward bias in the circuit under consideration because the collector battery always has a greater voltage than the base battery. This polarity serves to collect the electrons entering the collector region.)

6. Power amplification is obtained because the internal resistance of the collector circuit (the output circuit) greatly exceeds that of the input circuit. This permits the matched

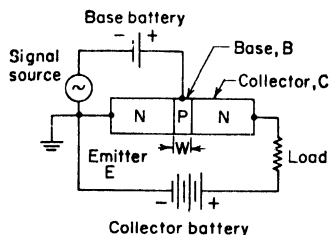
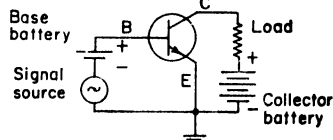



FIG. 10. Operation of transistor illustrated by case of n-p-n type in common-emitter circuit.

output resistance to be much larger than the transistor input resistance so that the output signal voltage can exceed the input voltage.

7. The reason for the difference of the input and output internal resistances is that, with respect to the base, the input is biased forward (giving low resistance) while the output is biased backward (giving high resistance).

For a more general explanation of transistor action it is necessary to recognize the existence of majority and minority current carriers. An n type of semiconductor has a surplus of free electrons, and these are the majority carriers for this type of semiconductor. The other type of carrier is the hole, or missing electron, which acts like a positive charge. A few of these are found in an n type of semiconductor and are the minority carriers. In the p type of semiconductor the roles are just reversed, the holes being the majority carriers and the electrons the minority carriers. The existence of the two types of carriers is well established, and measurements of their different mobilities and diffusion constants have been made. Fundamentally, the electron and hole types of carriers correspond to two different types of electron motion.

In terms of majority and minority carriers, the action of either an n-p-n or a p-n-p junction transistor can be summarized as consisting of the following three steps:

1. The emitter junction is biased forward, and the emitter injects its majority carriers into the base. These are minority carriers with respect to the base and are usually referred to in this manner.

2. These injected carriers largely pass on through the base, which is thin, and enter the collector region.

3. The collector is biased backward, giving a comparatively high internal resistance. This bias is opposite to that which would be required for a large original current at the base-collector junction, but is the polarity needed to collect and utilize the carriers entering the collector region from the base. These carriers are minority carriers in the base and are majority carriers in the collector region.

The ratio of collector current to emitter current is slightly less than unity, and is called the alpha (α) of the transistor. This is the signal-current ratio for the common-base connection. For the more frequently used common-emitter connection the corresponding quantity is the ratio of collector to base current, which is $\alpha/(1 - \alpha)$. Present transistors used as audio amplifiers typically afford alphas in the region of 0.97, corresponding to 32 for the common-emitter current gain.

Standard definitions of numerous transistor terms[1] and letter symbols[2] and standard test methods[3] have been adopted by IRE.

The standard letter symbols provide a systematic method of representing the various currents and voltages (average, rms, instantaneous total, instantaneous a-c component, etc.) and also the various impedances, admittances, and operating parameters.

One of the numerous differences between transistors and tubes is that certain transistor characteristics have a considerable dependence on temperature. This manifests itself in greater emitter current and lower collector resistance at higher temperatures, so that further rise of temperature is produced and there is a tendency toward overheating and damage. Stabilization arrangements have been devised,[4] the most widely used being a series emitter resistor.

There has been a steady reduction of the noise of transistors, but they do not yet ordinarily equal tubes in quietness. Recent transistor specifications include types with noise factors (Sec. 19) under 10 db.

At the higher frequencies the value of alpha falls off, and the frequency where it has declined to 70.7 per cent is defined as the *alpha cutoff frequency* and is listed in transistor specifications. Types for audio use typically have values of alpha cutoff in the region of 0.5 to 1 Mc. For subaudio power service this cutoff may be only 20 kc, while for i-f and r-f services and for fast computer use it may be many megacycles.

Progress in the design of transistors for higher frequencies or higher power levels has been steady, and further advances are confidently predicted.[5]

One widely used type of transistor for amplifier service is of the germanium-junction variety made by the *fused-junction* process, also called the *alloying* method. This consists of placing on each side of the base wafer a small amount of a doping agent which under heat will fuse to and alloy into the base material, thus forming the emitter and collector elements. For example, to make a *p-n-p* transistor of this kind, the base wafer is *n*-type germanium and small amounts of indium are placed on each side so as to produce the *p*-type germanium for the emitter and collector.

In distinction to this construction, the *grown-junction* types are formed by slowly pulling a forming crystal from a molten bath and changing the doping of the entire bath at appropriate times so that the emerging crystal has alternating *n* and *p* regions. This process has been extensively used for the production of *n-p-n* transistors.

Increased amplification has been obtained by increasing the emitter injection by the use of aluminum with indium as the *p*-type doping agent for *p-n-p* fused-junction transistors.[6]

Various methods have been used to increase the frequency range of transistors. An important cause of loss for the higher frequencies in junction transistors has been the finite time required for the minority carriers in the base to pass by diffusion from the emitter to the collector. Reduction of the width of the base region (*i.e.*, the distance from the emitter to the collector) has therefore been a primary aim. One method has been drilling a hole into the side of the base wafer to give a reduced dimension and forming the emitter and collector elements at this point.[7] This affords a base width of only 0.0005 in. An electrochemical etching spray is used in another process[8] to produce base widths of only 0.0001 to 0.0002 in. These transistors give

[1] "IRE Standards on Electron Devices: Definitions of Semiconductor Terms, 1954," Standard 54 IRE 7S2, *Proc. IRE*, **42**, 1505, October, 1954.

[2] "IRE Standards on Letter Symbols for Semiconductor Devices, 1956," Standard 56 IRE 28 S1, *Proc. IRE*, **44**, 934, July, 1956.

[3] "IRE Standards on Solid-state Devices: Methods of Testing Transistors, 1956," Standard 56 IRE 28S2, *Proc. IRE*, **44**, 1542, November, 1956.

[4] Wheeler, A. J., Thermistors Compensate Transistor Amplifiers, *Electronics*, **30**, 169, January, 1957.

[5] Kret, D. B., "A Report on the State of the Art of Semiconductor Devices," *Proc. of* 1956 *Electronic Components Symposium*, Engineering Publishers, Box 1151, New York, 1, N.Y.

[6] Armstrong, L. D., C. L. Carson, and M. Bentivegna, PNP Transistors Using High-emitter Alloy Materials, *RCA Rev.*, **17**, 37, March, 1956.

[7] Mueller, C. W., and J. I. Pankove, A PNP Triode Alloy Junction Transistor for Radio-frequency Amplification, *Proc. IRE*, **42**, 386, February, 1954.

[8] Bradley, W. E., *et al.*, The Surface-barrier Transistor, a series of five papers, *Proc. IRE*, **41**, 1702, December, 1953.

14 db gain per stage as video amplifiers with a 9-Mc band width and give reliable osciliator operation at 70 Mc.

Another technique for obtaining very thin layers is gaseous diffusion in a vacuum oven.[1] This permits thicknesses of less than 1 micron (10^{-6} m = 3.9×10^{-5} in.). Experimental *p-n-p* junction transistors made in this way have had alpha-cutoff frequencies of 500 Mc.

Silicon transistors are more difficult to produce because of the higher melting point of silicon compared with germanium. They have improved characteristics at higher temperatures and good high-frequency performance.[2]

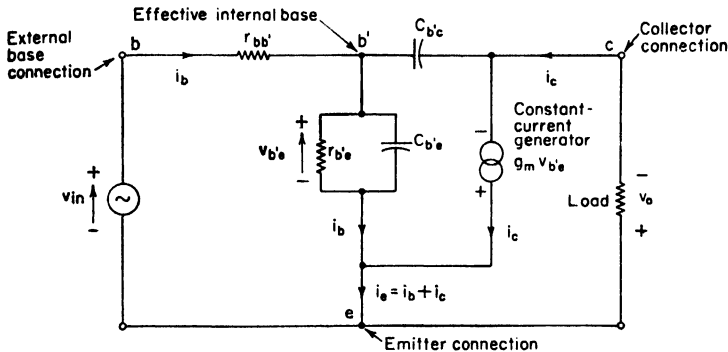



FIG. 11. Equivalent circuit of common-emitter-connected *p-n-p* or *n-p-n* junction transistor applicable over entire frequency range of operation.

An equivalent circuit representing a junction transistor of either the *p-n-p* or *n-p-n* type and applicable over the entire frequency range for which a particular transistor is useful has been devised independently by L. J. Giacoletto of RCA[3] and by C. R. Wilhelmsen of the Hazeltine staff.[4] This equivalent circuit is shown in Fig. 11 and exhibits a similarity to a pentode-tube circuit in that a voltage related to the input signal voltage actuates a constant-current generator in the output connections.

Resistance $r_{bb'}$ is the ordinary (or "ohmic") resistance of the base region between the external base connection and the effective internal base-emitter junction. Resistance $r_{b'e}$ is related to the forward-biased base-emitter junction, and capacitance $C_{b'e}$ is related to the diffusion or transit time of current carriers passing from the emitter through the base to the collector. The transconductance g_m is the ratio of collector-current increment to base-emitter voltage increment. Capacitance $C_{b'c}$ is the capacitance existing at the base-collector junction.

The value of $r_{b'e}$ in Fig. 11 is $KT/[qI_{dc}(1 - \alpha_0)]$, where K is the Boltzmann constant, T is the absolute temperature, q is the charge of an electron, I_{dc} is the emitter direct current, and α_0 is the low-frequency value of the collector-emitter current ratio. Substituting the known constants[5] and assuming room temperature, and also assuming that $\alpha_0 = 0.95$ and that the emitter current is 0.001 amp, it is found that $r_{b'e}$ has a value of 520 ohms. The diffusion capacitance $C_{b'e}$ is approximately $I_{dc}T_dq/(KT)$, where T_d is the average diffusion time of the current carriers in passing through the base region. For audio transistors T_d

[1] Lee, C. A., A High-frequency Diffused Base Germanium Transistor, *Bell System Tech. J.*, **35**, 23, January, 1956.

[2] Rittmann, A. D., and T. J. Miles, High-frequency Silicon Alloy Transistor, *Trans. IRE*, **ED-1**, 78, April, 1956. Tanenbaum, M., and D. E. Thomas, Diffused Emitter and Base Silicon Transistors, *Bell System Tech. J.*, **35**, 1, January, 1956.

[3] Giacoletto, L. J., Junction Transistor Characteristics at Low and Medium Frequencies, *Proc. Natl. Electronics Conf.*, **8**, 321, 1952.

[4] Wilhelmsen, C. R., Some Notes on the Hybrid-pi Transistor Equivalent Circuit, *Trans. IRE*, **BTR-4** (2), 92, March, 1958.

[5] To get the result in ohms, mks units should be used; in these terms $K = 1.37 \times 10^{-23}$ and

$$q = 1.60 \times 10^{-19}$$

may be ¼ μsec, giving a value of 0.01 μf for $C_{b'e}$. For higher-frequency transistors T_d may be only one-tenth as much, giving 0.001 μf. The transconductance g_m is $qI_{dc}\alpha_0/(KT)$, which for the same conditions is 37,000 μmhos. The feedback capacitance $C_{b'c}$ is dependent on the d-c voltage between collector and base, being MV_{CB}^{-n}, where M is a constant of proportionality and the exponent n is about one-half for alloy types and about one-third for grown-junction types. The value of $C_{b'c}$ is approximately equal to the published value of "collector capacitance." Typical values of the various elements of the equivalent circuit for representing two 1955 transistors are given in Table 8. One of these is a grown-junction *n-p-n* type offered for 455-kc service and the other is an alloy-junction *p-n-p* type offered for audio service.

Table 8. Typical Values of Elements of Equivalent Circuit of Fig. 11

Emitter d.c., amp	Grown junction, *n-p-n*				Alloy junction, *p-n-p*			
	$r_{bb'}$, ohms	$r_{b'e}$, ohms	$C_{b'e}$, μμf	g_m, μmhos	$r_{bb'}$, ohms	$r_{b'e}$, ohms	$C_{b'e}$, μμf	g_m, μmhos
0.0005	470	2,700	1,000	18,000	200	1,250	5,000	17,000
0.001	470	1,600	2,000	37,000	190	660	8,700	33,000
0.002	440	900	4,000	72,000	160	350	15,000	58,000

Values of Feedback Capacitance $C_{b'c}$

Collector-base d-c voltage	Grown junction, *n-p-n*, μμf	Alloy junction, *p-n-p*, μμf
1	11	76
2	9	57
3	7	35
4	5.6	25

38. Use of Transistors in Receivers. For audio amplification in receivers, as well as in other equipment,[1] transistors offer advantages, especially for compact designs such as for portable use. Transistor audio amplifiers are most frequently connected common-emitter because this connection affords the largest power gain. As a numerical example, the 2N190 transistor with a 9-volt battery in class A operation into a 12,000-ohm load will give a gain of 40 db (*i.e.*, a power ratio of 10,000) and deliver a maximum undistorted output of 0.002 watt. As a class B example, two 2N109s with a 9-volt battery and an 800-ohm load will deliver a maximum undistorted output of 0.2 watt at a gain of 29 db.

For low distortion the addition of negative feedback to the common-emitter circuit may be needed. Another means of obtaining high-fidelity operation is the use of the common-collector circuit, and good results have been reported with a 10-watt design of this kind using two 2N176s in push-pull in the driver and also in the output stage.[2]

Transistor audio circuits can often be designed with few or no transformers. If used, miniaturized designs are generally needed to avoid objectionable size and weight.[3]

[1] Shea, R. F., "Transistor Audio Amplifiers," Wiley, 1955. Trent, R. L., Design Principles of Junction Transistor Audio Amplifiers, *Trans. IRE*, **AU-3**, 143, September-October, 1955. Fewer, D. R., Design Principles for Junction Transistor Audio Power Amplifiers, *Trans. IRE*, **AU-3**, 183, November-December, 1955. Vallese, L. M., Optimum Design of Common Emitter Transistor Audio Amplifiers, *Proc. Natl. Electronics Conf.*, **11**, 919 (1955); also in *Communication and Electronics* (AIEE), No. 26, p. 391, September, 1956. See also Chaps. 10 and 15 of this handbook.

[2] Crow, R. P., and R. D. Mohler, Design of a High Fidelity 10 Watt Transistor Audio Amplifier, *IRE Convention Record*, **4**, Part 7, 142, 1956.

[3] Kajihara, H. H., Miniaturized Audio Transformer Design for Transistor Applications, *Trans. IRE*, **AU-4** (1), January-February, 1956.

Noise originating in transistors departs from the well-known random or "white" fluctuation noise (described in Sec. 27) in the direction of greater noise at lower frequencies; in fact, transistor noise power is inversely proportional to the frequency. Despite this fact, it is possible to obtain satisfactory operation of transistor audio amplifiers with respect to noise by reasonable care in the choice of the type of transistor according to noise-factor ratings and the requirements of the particular application. Types rated below 10 db are available, although as of 1957 most types were higher.

Noise factors as low as 2 db have been obtained in experimental *p-n-p* alloy-junction transistors,[1] and steady improvement in commercial types is to be expected.

The existence of both the *p-n-p* and the *n-p-n* types of transistors permits the cooperative use of the two types in circuit in a manner which is not possible with tubes. This is called *complementary symmetry.*[2] If complementary transistor types can be obtained which have equal corresponding characteristics within the necessary close tolerances, this circuit arrangement has the advantage that a push-pull stage can be designed without transformers and without any input phase inversion. In principle, complementary symmetry can also be applied to a two-stage amplifier with a single transistor in each stage.

For 455-kc i-f service, transistors with an alpha cutoff of 5 Mc or higher will typically afford power gains of 30 db or more per stage. The common-emitter connection is normally used with coupling circuits to match the transistor input, which for this connection is 300 to 1,000 ohms, and also match the transistor output of 20,000 to 100,000 ohms.[3] These types of transistors have collector feedback capacitances ($C_{b'c}$ in Fig. 11) of 5 to 15 $\mu\mu$f and base resistances ($r_{bb'}$) of 75 to a few hundred ohms.

Transformers and other coupling networks for transistors are quite different from those in tube practice because of the different impedance levels. The difference of impedance is especially important on the input side, *i.e.*, at the base-to-emitter input terminals. Suitable design procedures for these coupling systems are available in the literature.[4]

The substantial value of the collector feedback capacitance makes necessary the consideration of neutralization for operation at 455 kc and above. Whether or not to provide neutralization for a 455-kc amplifier is one of the questions to be answered in the design. With common-emitter operation partial neutralization is easily obtained by the addition of neutralizing capacitors from base to base of adjoining stages. More refined methods, sometimes called *unilateralization,* afford virtually complete removal of feedback.[5]

Automatic gain control can be applied to transistors by utilizing the dependence of the gain on either the emitter d.c. or the collector d-c voltage. Decrease of either of these quantities reduces the gain.[6] The addition of a biased diode as another means of control at high input levels may also be advisable.

The small amounts of d-c power required for transistor operation have led to suggestions of equipment utilizing only "free" power, such as solar energy,[7] the acoustic energy of the voice, rectification of waves received from nearby strong radio stations,

[1] Englund, J. W., Noise Considerations for P-N-P Junction Transistors, "Transistors I," RCA Laboratories, Princeton, N.J., 1956; see pp. 309–321.

[2] Stanley, T. O., and T. M. Scott, Design Considerations in Class-B Complementary Symmetry Circuits, "Transistors—I," RCA Laboratories, Princeton, N.J., 1956; see pp. 497–514.

[3] Cheng, C. C., Simplified Design Procedures for Tuned Transistor Amplifiers, *RCA Rev.*, **16,** 339, September, 1955.

[4] Hellstrom, M. J., Design of Double Tuned IF Transformers for Transistor Amplifiers, *IRE Convention Record*, **4,** Part 3, 69, 1956. Webster, R. R., How to Design I-F Transistor Transformers, *Electronics*, **28,** 156, August, 1955. Rand, R. C., and J. B. Oakes, Single and Double Tuned Transistor IF Amplifiers, *Proc. Natl. Electronics Conf.*, **11,** 52, 1955. Savant, C. J., and C. A. Savant, Transistor Transformer Design Nomographs, *Electronics*, **30,** 204, Mar. 1, 1957; applies to a-f amplifiers.

[5] Holmes, D. D., and T. O. Stanley, Stability Considerations in Transistor I-F Amplifiers, "Transistors—I," RCA Laboratories, Princeton, N.J., 1956; see pp. 403–421.

[6] Chow, W. F., and A. P. Stern, Automatic Gain Control of Transistor Amplifiers, *Trans. IRE*, BTR-1 (2), 1, April, 1955.

[7] Raisbeck, G., The Solar Battery, *Sci. American*, December, 1955.

the operation of a small hand lever, etc.[1] Silicon transistors are of special interest in this connection.[2]

An all-tube vehicular receiver for military service, operating over the range from 0.5 to 32 Mc, was advantageously redesigned by introducing transistors for all circuit functions except in the front end. The transistorized circuits included the 455-kc i-f amplifier with choice of three band widths, a.v.c., squelch, noise limiter, beat-frequency oscillator for c-w telegraph reception, and a frequency calibrator.[3]

An analysis of transistorization in air-borne equipment revealed a 20 per cent saving in weight, 25 per cent saving in physical volume, and improved reliability of operation.[4]

Broadcast receivers employing transistors,[5] with all power supplied from one or more batteries, can be designed in a variety of sizes, depending chiefly on the required performance and battery life and on the size of a satisfactory miniature loud-speaker.

For medium sizes, such as with 9 in. as the maximum dimension and 4 lb as typical weight, the circuit can include six or seven transistors and a crystal-diode detector, the transistors serving as converter (with or without separate oscillator), two-stage i-f amplifier, first audio stage, and push-pull second audio stage.[6] A paper presented in 1955 by J. L. Nygaard of the Texas Instrument staff reported that a sensitivity of 80 db below 1 volt per m for 0.05 watt output is obtainable in a design totaling six transistors and two diodes and having the following features: (1) a ferrite antenna core of 0.35 in. diameter and 8 in. length; (2) transistor converter; (3) two 2N146 transistors as neutralized i-f stages; (4) crystal-diode detector; (5) transistor first audio stage; (6) two transistors as push-pull second audio stage; (7) crystal antioverload diode connected between ground and a tap 33 per cent up on the antenna winding; and (8) a 9-volt battery. With the antenna-circuit diode and a-g-c bias applied on only the first i-f stage, an a-g-c figure of merit of approximately 60 db was obtained. In general, the performance is similar to that of the conventional four-tube-plus-rectifier receiver in which the four tubes serve as (1) converter; (2) i-f amplifier; (3) detector, a.v.c., and first audio stage; and (4) output audio stage.

For a smaller design, such as with a maximum dimension of 5 or 6 in. and weight of 1 lb, the design can incorporate five to seven transistors and may also include a crystal-diode detector. Two of the transistors serve as i-f stages and one as audio stage. This may be the only audio amplification, or a push-pull output stage may also be provided. A single-transistor converter may be included or separate oscillator and mixer. The performance will vary widely depending on the number of stages. The i-f stages can often advantageously be neutralized, and a.v.c. is normally applied to the first one. Typical battery life is 100 hr.[7]

Broadcast-band receivers for automobiles with 12-volt batteries can be designed with a power transistor as the output stage and with tubes, specially offered for 12-volt plate supply, in the other stages. In this way vibrator, power transformer, and rectifier are eliminated. Transistors offered for such service include the 2N155, 2N176, 2N178, 2M242, 2N257, and the 2N268; these furnish substantial outputs, such as 2.5 watts with 5 per cent total harmonic distortion, as required for automobile service. Data on tubes rated at 12 volts plate supply are given with further treatment of automobile receivers in Sec. 153 of this chapter.

Subject to the availability of suitable types of transistors, automobile receivers can be designed on an all-transistor basis. Compared with an all-tube receiver, this affords a 90 per cent saving in operating power and a substantial reduction in receiver size. An experimental design of this kind[8] with r-f stage, mixer, separate oscillator, two 455-kc i-f

[1] Hollmann, H. E., Designing "Free-power" Transistor Apparatus, *Tele-Tech and Electronic Industries*, **15**, 66, August, 1956. Hollmann, H. E., Designing "Free Power" AM and FM Transistorized Receivers, *Tele-Tech and Electronic Industries*, **15**, 54, September, 1956.

[2] Keonjian, E., Micro-power Operation of Silicon Transistors, *Tele-Tech and Electronic Industries*, **15**, 76, May, 1956.

[3] Schwartz, S., Transistorized Receiver for Vehicular Radio, *Electronics*, **29**, 217, October, 1956; also *Trans. IRE*, PGVC-7, 41, December, 1956.

[4] Steutzer, O. M., Transistors in Airborne Equipment, *Proc. IRE*, **40**, 1529, November, 1952.

[5] 1956 Transistor Portable Design, *Tele-Tech and Electronic Industries*, **15**, 100, April, 1956; schematics of seven manufactured models are given.

[6] Sheehan, W. E., and J. H. Ivers, Design of Transistorized High-gain Portable, *Electronics*, **28**, 159, March, 1955.

[7] Holmes, D. D., T. O. Stanley, and L. A. Freedman, A Developmental Pocket-size Broadcast Receiver Employing Transistors, *Proc. IRE*, **43**, 662, June, 1955.

[8] Freedman, L. A., T. O. Stanley, and D. D. Holmes, An Experimental Automobile Receiver Employing Transistors, *Proc. IRE*, **43**, 671, June, 1955.

stages, transistor detector, single first audio, and push-pull second audio totaled nine transistors. This design had a performance comparable with conventional all-tube receivers, including a sensitivity (measured at 1 watt output) of 114 db below 1 volt, an a-v-c figure of merit of 63 db, and an undistorted power output of 2 watts.

Transistors are a major element in the miniaturization of receiving equipment, which is discussed in Sec. 130.

R-F AMPLIFIERS[1]

39. Inclusion or Omission of R-f Amplifier. The majority of home broadcast receivers now in use are two-gang a-c-d-c superheterodynes with four tubes and power rectifier, as this design has been found the most popular combination of low price and acceptable performance. A set of this kind has no r-f stage, the four tubes being (1) the converter, (2) the i-f stage, (3) the diode detector and first audio stage in one envelope, and (4) the output stage. The sections of the two-gang variable capacitor tune the antenna coupling and the oscillator section of the converter.

Important advantages afforded by the use of an r-f amplifier with three-gang tuning include prevention of radiation of the oscillator output, reduction of the effect of the high tube noise of the converter, and a large improvement in the rejection of interference at the image and intermediate frequencies.

In special-purpose receivers, such as used in amateur, police, and point-to-point services, one or two tuned radio stages are generally provided. For frequencies up to somewhat in the v-h-f band (*i.e.*, above 30 Mc), pentodes are suitable for this service, especially the 6AK5, 6CB6, and 6DC6. However, at higher frequencies a point is reached where the tube noise of the pentode is objectionable and other methods are necessary to obtain sufficiently low noise factor.[2] Triodes with grounded-grid operation are suitable, especially if preceded by a grounded-cathode triode in the widely used *cascode* circuit.[3] Tubes offered for such use include the 6BQ7A, 6BZ7, and 6AJ4. For still higher frequencies the traveling-wave tube holds promise of affording low-noise amplification. For further treatment of receivers operating above 30 Mc see the text beginning at Sec. 154 below.

An untuned r-f stage used occasionally in medium-sized chassis for home receivers affords the advantages of suppressing oscillator radiation and giving some gain, at the cost of one tube and a very few inexpensive additional parts. Resistance coupling is generally used in such stages, giving a stage gain of 10 to 13 db (3.2 to 4.4 times voltage gain) in the m-f band. If h-f international broadcasts are to be received, a high-transconductance tube must be used in the r-f stage to obtain appreciable gain, and it may be advisable to insert a series coil to resonate with the capacitances of the two tubes at a frequency near the upper limit of the range to be received. For example, a coupling to go up to 16 Mc and work between 10-$\mu\mu$f tube capacitances may include a coil to resonate at this frequency with the 5-$\mu\mu$f series equivalent of the two capacitances. This requires an inductance of 40 μh. A coil of medium quality is sufficient because high Q will give too sharp a peak at the higher frequencies of the desired tuning range.

A series trap for rejection of i.f. may be considered as an improvement for an untuned stage. This has the additional advantage of markedly improving the stability, *i.e.*, freedom from regeneration and tendency to oscillate, of the i-f amplifier of the receiver. Typical trap constants for 455 kc are in the neighborhood of 1,000 μh and 125 $\mu\mu$f, or 2,000 μh and 62 $\mu\mu$f. Factory adjustment and provision for field adjustment are needed. A high Q for this coil is desirable.

40. Inductive Tuning. Automobile broadcast receivers very often employ inductive tuning of the type in which a movable core of compressed powdered iron is inserted varying distances into the coil form to give the required inductance. Such cores are

[1] Sturley, K. R., "Radio Receiver Design," Part I, 2d ed., Wiley, 1953; see pp. 214-289. Langford-Smith, F., "Radiotron Designer's Handbook," 4th ed., RCA, Harrison, N.J., 1952; see pp. 922-946.

[2] Moxon, L. A., "Recent Advances in Radio Receivers," Cambridge, 1949; see pp. 20-53.

[3] Wallman, H., A. B. MacNee, and C. P. Gadsden, A Low-noise Amplifier, *Proc. IRE*, **36**, 700, June, 1948.

called "slugs." A typical interstage coupling of this kind has an almost constant Q with a value of about 80 over the band. The inductance range must be about 10 to 1 to cover the range 1,600 to 550 kc. A typical inductance range is 30 to 300 μh resonating with a fixed capacitor of about 250 μμf.

41. Unicontrol Tracking.[1] The superheterodyne with an r-f stage normally requires the alignment of three tuned circuits of quite different characteristics. The r-f interstage coupling is taken as reference, in comparison with which the antenna coupling must accommodate a fixed loop antenna or an open antenna usually having any capacitance within a wide range. The oscillator circuit must, of course, operate at a frequency lying above or below the signal frequency by an interval equal to the i.f. Procedures for proportioning the oscillator circuit are given in Secs. 58 to 60 of the present chapter. For push-button tuning refer to Sec. 61 *et seq.*

In the m-f broadcast band the usual tuning reactors for an r-f stage with capacitive tuning are a coil of about 200 μh and a variable capacitor with a maximum of about 420 μμf. With inductive tuning, typical values are a fixed capacitor of about 250 μμf and a moving-core coil with inductance range from 30 to 300 μh. The capacitors in inductive tuning are normally not actually fixed but are adjustable for alignment purposes.

42. Cross Modulation in R-f Tubes. When a receiver is tuned to a strong signal and there is also present another strong signal of not too different frequency, cross modulation or cross talk may occur. During intervals when there is no modulation on the desired carrier, the modulation of the undesired carrier is heard, but if the desired carrier is absent, the undesired program disappears. The difficulty may occur in an r-f stage in either a superheterodyne or a t-r-f receiver. The cause is the operation of the tube with a total input so large that the linear part of its characteristic is exceeded. Under these circumstances, the tube acts as a modulator and produces a variation in the amplitude of the desired signal in accordance with the modulation of the undesired signal; *i.e.*, the tube gives to the desired signal a modulation from the undesired signal, whence the name "cross modulation." Since the interference becomes part of the modulation of the desired carrier, no subsequent operations can remedy the condition. No special frequency relation exists between the desired and undesired carrier frequencies. In this respect the phenomenon differs from external cross modulation (Sec. 30) and from spurious superheterodyne responses (Sec. 47).

To avoid this fault, the gradual-cutoff, or "variable-mu," type of tube has come into use. Such a tube, if preceded by a single tuned circuit, is normally sufficient to prevent difficulty on this account.

Special care to reduce cross modulation is needed in the design of receivers for mobile reception of radiotelephone messages at 50 and 160 Mc. The selectivity measured with the usual one-signal method may be far less than the effective selectivity, which is impaired by the cross modulation and revealed by the two-signal selectivity test.[2]

The degree to which a given tube will produce this interference can be measured by noting the amount of increase of transconductance when the tube is measured at a low level of signal and then at a substantial level.[3] Another method, described in the same reference, is to measure the tube with a low value of signal and then superpose a substantial signal of considerably higher frequency, again noting the increase of transconductance. The greater the increase of transconductance, the more cross modulation the tube will produce.

SUPERHETERODYNE PRINCIPLE

43. Basic Advantages. Practically all receivers now being made are superheterodynes. This applies to the wide range from low-price broadcast receivers to excellent special-purpose designs.

[1] Langford-Smith, F., "Radiotron Designer's Handbook," 4th ed., RCA, Harrison, N.J., 1952; see pp. 1002–1019.

[2] Manke, A. G., Effect of Front End Receiver Design on Over-all Performance, *Trans. IRE*, PGVC-5, 53, June, 1955.

[3] Herold, E. W., Simple Method for Checking R-f Distortion or Cross Modulation of Pentode Amplifier Tubes, *Electronics*, April, 1940, p. 82.

The advantages of the superheterodyne are primarily the advantages of i-f stages in comparison with t-r-f stages. Much greater amplification and selectivity per stage are realized at i.f. (1) because no provisions have to be made in the i-f amplifier for variable tuning for the selection of stations, and (2) especially because the value of the i.f. can be chosen for the best performance. An additional advantage of the superheterodyne is the increase of stability afforded by the use of an additional frequency range; gains that would almost surely cause oscillation in a t-r-f receiver are easily obtained with the superheterodyne.

44. Analysis of Typical Superheterodyne Performance. The results normally obtained with a good grade of superheterodyne broadcast receiver having two i-f stages are shown in Figs. 12 and 13 in terms of total voltage gain and selectivity. The performance of any one stage is shown by the ratios of the ordinates of the proper two curves at the various frequencies. For example, in Fig. 12 the voltage gain in the second i-f stage at 1,000 kc is total gain to the second detector divided by the total gain to the second i-f grid, or 960,000/16,000, which is 60 times or 36 db. In a similar way the selectivity of each stage at any frequency is obtainable from Fig. 13.

Typical over-all performance of broadcast receivers is given in Sec. 23.

FIG. 12. Voltage gain in successive stages of superheterodyne.

45. I.F. Higher than Signal. In a few cases it is desirable to have the i.f. higher than the signal, *i.e.*, to raise the incoming signal in frequency rather than lower it. Such a circuit reduces the wavelength of the received signal and therefore has been called an "infradyne."

An instance is the 500-kc autoalarm set[1] of the Radiomarine Corporation of America where good image rejection is obtained by the use of a 1,100-kc i.f. The oscillator operates at 1,600 kc, and the image is at 2,700 kc. An autoalarm set maintains a watch for the international distress signal at ship and land stations when the operator is off duty or doing other work.

The use of an i.f. above the signal range gives excellent image rejection but is subject to spurious response at subharmonics of the i.f. and in the case of two signals whose sum equals the i.f. In the former case the converter produces a fixed-frequency i.f. which will "beat" with the normal i.f. (whose exact value varies as the tuning dial

[1] Byrnes, I. F., and H. B. Martin, Automatic Alarm, *RCA Rev.*, January, 1937, p. 49.

is adjusted) and thereby produce a whistle, or "tweet." In the second case the two signals also produce a fixed i.f. with similar results.

46. Double Superheterodynes. For special purposes there is often a need for repeating the frequency change an additional time. This is especially the case in the v-h-f range. In a double superheterodyne there are two i.f.'s, the first of which is usually the higher and is chosen to give the required image rejection. The second i.f. is then chosen for best amplification and close-in selectivity considering the band width necessary to handle the signal. The first i.f. is sometimes adjustable as an additional means of station selection. Since a double superheterodyne has two frequency converters and a detector, it is sometimes called a "triple-detection" receiver.

FIG. 13. Increasing selectivity in successive stages of a superheterodyne receiver.

The value of the first i.f. in double superhets usually lies between 1.5 and 50 Mc for services such as single and multiplex phone. (For radar, where a very wide band width is necessary, the single superheterodyne is generally used.) The second i.f. ranges ordinarily from 455 to 5,000 kc.

The disadvantages of the double superhet are (1) the added complexity and (2) the tendency to give numerous and prominent spurious responses. Considerable design and development work are therefore generally required to produce a model with acceptable performance.

One distinct type of double superheterodyne uses the same oscillator frequency for both conversions. As an instance, the model AN/TRC-1 series of f-m radiotelephone equipments[1] for a range of 70 to 100 Mc have a double-superhet receiver with a single oscillator whose frequency can be adjusted from 37.5 to 52.5 Mc. The first i.f. is variable in all single-oscillator double superheterodynes if a tuning range is provided, and in the present instance it may have any value from 32.5 to 47.5 Mc. The second i.f. is fixed at 5 Mc. Assume that a signal of 80 Mc is to be received, for which the oscillator is set at 42.5, the first i.f. is 37.5, and the second i.f. has its constant value of 5 Mc. As another example, for receiving a 100-Mc signal the oscillator is at 52.5 and the first i.f. is 47.5 Mc. If f is the signal frequency to be received and F is the fixed second i.f., then the oscillator frequency f_o and the first i.f. f_1 are as follows:

$$f_o = \frac{(f + F)}{2} \quad \text{and} \quad f_1 = \frac{(f - F)}{2}$$

[1] Marks, W. S., Jr., O. D. Perkins, and W. R. Clark, Radio-relay Communication Systems in the United States Army, *Proc. IRE*, August, 1945, p. 502.

These formulas give oscillator frequencies higher than the first i.f.'s; if oscillator frequencies lower than the first i.f.'s are desired, the formula for f_1 can be used to get the oscillator frequency, and the formula for f_o to get the i.f.

A multiple superheterodyne of special interest was used on Jan. 10, 1946, by Lt. Col. John H. DeWitt, Jr., and four assistants of the Signal Corps to receive the first radar echo from the moon.[1] The achievement took place at the Evans Signal Laboratory, Belmar, N.J. To obtain the very weak echoes of 0.002 to 0.02 μv per m at the antenna, corresponding to 0.01 to 0.1 μvolt delivered to the preamplifier, it was necessary to reduce the effect of noise by greatly restricting the band width and providing a receiver of very low noise figure. The equipment consisted of a much modified SCR-271 radar in which the original operating frequency of 111.5 Mc was retained. A band width of only 57 cycles was used, which was sufficient for the pulse of ¼ sec duration sent out every 4 sec. The trip to and from the moon requires a total of 2.4 sec. To realize the necessary narrow band width, very precise control of frequency was required, and in addition the Doppler change of frequency on reflection at the moon had to be taken into account. This effect amounted to 300 cycles and changed in sign according to whether the observations were being made at moonrise or moonset. The most practicable way of meeting the requirements as to frequency stability and band width with the available equipment was to employ a quadruple superheterodyne, the i.f.'s being approximately 33 Mc, 6.7 Mc, 1.6 Mc, and 180 cycles. The 57-cycle band width was obtained with a pass band of approximately 150 to 210 cycles for the last i-f amplifier centering at 180 cycles. The echo from the moon was seen on a scope and also heard as a pulse of 180-cycle tone every 4 sec.

The antenna had a gain of 23 db over an isotropic radiator. The preamplifier of the receiver was a special design with a noise figure of only 5 db. It had three stages of which the first two were grounded-grid 6J4 triodes, a miniature type with the high transconductance of 12,000 μmhos. The noise power was doubled by the acceptance of the image in the frequency conversion from 1.6 Mc to 180 cycles, so that the noise figure of the entire receiver was 8 db.

47. Spurious Responses in Superheterodynes. The selectivity of a superhet in terms of selectances and of band widths at certain input ratios is not a complete indication of the actual selectivity under all conditions, as this type of receiver is susceptible to certain kinds of interference which are not encountered with a t-r-f receiver.[2] The susceptibility to these interferences is a result of frequency conversion. Considerable care may be necessary to distinguish interference of this general type from that due to cross modulation or external cross modulation.[3]

The discussion in the following paragraphs is primarily for the case where the i.f. is lower than the received frequencies. For spurious responses when the i.f. is above the received frequency range, see Sec. 45.

48. Image Interference. The most important spurious response in a superheterodyne is usually the image. With the set tuned to a particular desired signal, the image lies on the other side of the oscillator, each being separated from the oscillator by the i.f. For example, a receiver tuned to 1,000 kc and having an i.f. of 455 kc is subject to image interference at 1,910 kc. Any receiver with this value of i.f., when tuned to any station, is subject to image interference at a frequency 910 kc higher than the desired station.

Since the function of the converter is to produce the difference between applied frequencies, it cannot distinguish between the signal and the image and produces i-f output from each. The only remedy for image interference is increased general r-f selectivity ahead of the converter, or special circuits[4] designed to have a high attenuation at the image frequency, wherever it may be, as the tuning dial is adjusted to select any desired station. The simplest means for obtaining image rejection is to tap down the output lead on the tuning inductor. This can be done to give theoretically infinite attenuation at one frequency and substantial improvement over a band. The tap

[1] Mofenson, Jack, Radar Echoes from the Moon, *Electronics*, April, 1946, pp. 92–98.

[2] Morgan, Howard K., Interfering Responses in Superheterodynes, *Proc. IRE*, October, 1935, pp. 1164–1170.

[3] A useful comprehensive table, "Analysis of Radio Interference Phenomena," appeared in *RCA Radio Service News*, February, 1946, and was reprinted in *Radio News*, June, 1946, p. 54.

[4] Wheeler, Harold A., Image Suppression in Superheterodyne Receivers, *Proc. IRE*, June, 1935, p. 569.

an antenna dimension can be expressed in wavelengths by the ratio of the physical length to the wavelength (expressed in same units as length) corresponding to the frequency, or in electrical degrees (or radians) by the product of the dimension in wavelengths and 360 deg (or 2π radians). For example, at a frequency of 10 kc, *i.e.*, 3×10^4 m wavelength, the vertical length of an antenna 300 m high can be expressed as $V = 300 \text{ m} = 0.01\lambda = 3.6 \text{ deg} = 0.0628$ radian.

The lowest frequency at which the antenna system in Fig. 4 is resonant is defined as the fundamental frequency, and the corresponding free-space wavelength is defined as the fundamental wavelength λ_0. The effective (or apparent) electrical length G of this antenna system at the fundamental frequency is $\lambda_0/4$, or 90 electrical degrees. For any vertical length V, the effective length is a function of the length of the top-loading section H and the diameter of the vertical section D.

FIG. 4. Empirical values for the constant K used in determining the wavelength of T antennas.

If the fundamental frequency, and hence wavelength, of this antenna system is known, its effective electrical length G at any other frequency can be obtained from

$$\begin{aligned} G &= \tfrac{1}{4}(\lambda_0/\lambda) && \text{wavelengths} \\ &= 90(\lambda_0/\lambda) && \text{degrees} \\ &= (\pi/2)(\lambda_0/\lambda) && \text{radians} \end{aligned}$$

where λ is the wavelength corresponding to the frequency.

The fundamental wavelength can be obtained from the following approximate relations or from measurements on scale models. For an antenna without top loading, or a vertical tower of uniform cross section, the following formula applies to within approximately 4 per cent.

$$\lambda_0 \doteq 1.33V \qquad \text{m}$$

where V = vertical length (height), ft.

For top-loaded symmetrical T antennas the fundamental wavelength can be predicted only approximately from simple formulas, and for large-area flattop systems it usually must be estimated or measured from scale models. Many empirical factors are the cause of this uncertainty, including the ratio of horizontal to vertical lengths, the capacitance of insulators, the proximity of grounded or insulated towers, the number of conductors used and their cross-sectional configuration, reflections from the T junction, and perhaps other factors. It has been found from practical experience that the following empirical relation for the fundamental wavelength can be used for a symmetrical T antenna.

$$\lambda_0 \doteq K(V + H/2) \qquad \text{m}$$

where the antenna dimensions are in feet, λ_0 in meters, and K is taken from Fig. 4. This figure was derived from 24 actual antennas which were carefully measured. The antennas were of different sizes and with several different tower arrangements. This unusual curve accounts within 5 per cent for all but four of the antennas. The corners in the curve are based on reliable data and so were not rounded out. The value at the right-hand extremity of the curve is based on three identical l-f antennas constructed in three different locations and measured by separate qualified individuals with exact concordance.

In some cases the effective electrical length G of an inverted L or T antenna can be taken as the sum of the vertical and horizontal parts in electrical degrees (or radians) provided that the horizontal distances from the junction with the vertical are electrically small. For capacitance-area loading of empirical form, calculations may be difficult. In such cases, it is desirable to use a scale model (the scale being in size

directly proportional to the wavelength) erected over a metallic plane of area very large with respect to the flattop. The supports should also be modeled and the location of antenna and guy insulators simulated. For reactance measurements, the cross section of the antenna conductors should also be to scale.

9. Radiation Resistance and Field Strength. The radiation resistance can be increased by increasing the vertical length (see Fig. 5) and/or by producing, as nearly

FIG. 5. (*Left*) Radiation resistance for electrically short antennas vs. height of vertical portion (degrees) and (*right*) ratio of currents at top and base of vertical portion.

as possible, a uniform current distribution on the vertical portion of the antenna. The latter is usually done by some capacitance loading at the upper end in the form of an inverted L, a T, or a large flattop in the form of a ring or spider.

The radiation resistance of a loaded antenna can be calculated from the following (see Fig. 5).

$$R_r = \frac{30}{\sin^2 (L + B)} \left[\frac{4L^2}{3} \sin^2 B + \frac{2L^3}{3} \sin 2B + \left(\frac{17}{30} \cos 2B - \frac{7}{30} \right) L^4 \right] \quad \text{ohms}$$

where L = vertical length (height), deg

B = effective lengthening of the antenna due to the top loading, deg; $L + B = G$

When $B = 0$, $R_r = 10L^2$ (L in radians). The above relationships are valid for $L \leqq \lambda/12$ ($L \leqq 30$ deg) and are shown in Fig. 5.

FIG. 14. Chart for identifying whistles heard with superheterodynes having an intermediate frequency of 455 kc.

where m = desired modulation percentage
E_w = output whistle voltage
E_s = signal output for 30 per cent modulation at 400 cycles

As an example, if E_w should equal E_s, the value of m would be 30 per cent, which would be reasonable as indicating equality to the 30 per cent signal modulation.

The test is repeated at representative signal levels. The results give a clear picture of the minimum whistle interference which characterizes the set, since in this test there is no interference to aggravate the situation. In this test, whistles will be found at harmonics of the i.f. and possibly at other points.

A two-signal test is also provided in the IRE Standards which simulates the practical case where the receiver is tuned to a given frequency but subject to interference over a wide frequency range. One of the signal generators supplies the input to which the set is tuned, and the other generator is used to explore for whistles. As each one is located, a measurement is made of the input of the "interfering" signal generator necessary for the whistle, when tuned to 400 cycles, to be 30 db below standard output. The results are plotted as a spectrum with frequencies as abscissas. The ordinates are decibels below 1 volt plotted linearly or microvolts plotted logarithmically.

FREQUENCY CONVERTERS

53. Essential Function. The frequency-changing function is performed by a modulator, or mixer. This requires the r-f signal and also the output of a local oscillator, and delivers an i-f signal. The general theory of modulators is given in Chap. 12. The modulator, or mixer, and the oscillator are generally called a "converter." The mixer and oscillator of a converter may be in either the same or separate tubes. One classification of converters is according to whether the signal and heterodyne inputs are applied at the same electrode or at separate electrodes and, in the latter case, according to whether the electron stream is first influenced by the heterodyne or by the signal voltage. The mixer function can be performed by any element which can be so operated as to have a nonlinear instantaneous input-output characteristic, including tubes of diode, triode, and multielement types, and also crystal diodes and transistors.

54. Converter Performance in M-f Broadcast Band. The most widely used converter tubes for the reception of sound broadcasts are heptode oscillator-modulators, such as the 6BE6 and the 12BE6, having the oscillator function associated with the first grid and the signal input voltage applied on a later grid. These are also called *pentagrid converters.* They afford conversion transconductance (ratio of i-f current to applied r-f signal voltage) of 425 to 550 μmhos with a plate supply of 250 volts. The corresponding battery tube, the 1R5, is rated at 235 to 330 μmhos at plate voltages from 45 to 90 volts. In socket-power operation, "conversion gain" (i-f voltage on first i-f grid divided by r-f signal voltage on converter input grid) of forty to sixty times is normally obtained over the standard m-f broadcast band. A-v-c bias voltage can be applied to the return connection of the r-f signal input grid, and lower conversion gain can thus be obtained for improving the a-v-c figure of merit.[1]

55. Converters for the 100-Mc F-m Band. A frequency converter for the f-m band of 88 to 108 Mc is available in the 6BA7 heptode, which for separate-oscillator service is rated at 900 to 950 μmhos conversion transconductance. It can also be used in a self-oscillating design.

For high-grade performance, such as in tuners for high-fidelity installations, separate triode oscillators are normally provided, often in the same tube envelope with the mixer. Examples of such triode-pentodes are the 6U8 and the 6X8. The 12AT7 double triode affords separate triode sections suitable for oscillator and mixer use. If a.f.c. is to be provided, the mixer can be a separate triode (*e.g.*, 6AB4) or pentode (*e.g.*, 6SH7, 6BC5, or 6AU6), and the oscillator and a-f-c functions can be obtained in a double triode, such as the 12AT7, 6J6, or 6BK7A.

Oscillator frequency drift accompanying warmup of the receiver can be controlled by a.f.c. (per Sec. 96 *et seq.*) or by careful attention to the characteristics of the oscillator tube, the various other circuit parts, and the use of temperature-compensating capacitors.[2]

[1] Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 962–1001. van der Zeil, A., "Noise," Prentice-Hall, 1954; see pp. 233–261.

[2] Pan, W. Y., Frequency Characteristics of Local Oscillators, *RCA Rev.*, **16**, 379, September, 1955.

56. Germanium Crystal Converter for V.H.F. At frequencies in the v-h-f and lower part of the u-h-f range germanium crystals offer the advantages of compactness, absence of heater power requirements, and good electrical characteristics. The characteristics of four representative types are given in Table 9. Numerous other types of germanium diodes having EIA designations are available with substantial ranges of characteristics and in various housings. Frequencies up to 900 Mc are within the ratings of the 1N124A, 1N133, and 1N147, which are offered for mixer service in u-h-f television receivers. The 1N93 and 1N153 are rated to 300 volts instantaneous inverse potential. The 1N56A and 1N71 can handle 0.060 amp average current and have minimum forward current at +1 volt of 0.015 amp.[1]

Table 9. Characteristics of Typical General-purpose Germanium Crystal Diodes

Property	1N34	1N126	1N38	1N127
Max working volts	60	60	100	100
Min forward current at 1 volt, amp	0.0085	0.005	0.0075	0.003
Max reverse current	800 μa at −50 volts	850 μa at −50 volts*	25 μa at −10 volts	25 μa at −10 volts†
Max transient or surge current, amp	0.5	0.35	0.5	0.3
Max continuous d.c., amp	0.04	0.03	0.0225	0.03

* Maximum reverse current at −10 volts is 50 μa.
† Maximum reverse current at −50 volts is 300 μa.

57. Silicon Crystal Converter for U.H.F. and S.H.F. In the range from 1,000 to 55,000 Mc, the fixed silicon crystal with tungsten point is generally used as a converter because of its low noise and moderate conversion loss.[2] Below 2,500 Mc tube converters can be employed with the advantage of greater electrical ruggedness but the drawback of poorer operating performance. Table 10 gives characteristics of various silicon crystals which are available for different frequencies.

Crystals must be protected from electrical overload, such as due to poor duplexer operation in radars. The burnout-test energy in a "spike" of admitted transmitter output is only 0.3 to 2 ergs for 3,000-Mc crystals, 0.3 to 1 erg for 9,000- and 10,000-Mc types, and 0.1 erg for the 25,000-Mc type 1N26. For 1,000-Mc service, the Western Electric 1N25 is available with the high value of 6.5 watts burnout pulse power.

At lower frequencies crystals produce more noise for a given band width.[3] This is true below 1 Mc and especially in the a-f range.

Most types of silicon crystals are designed for operation at or close to a particular radio frequency. In distinction, the 1N286, 1N358, 1N369, and 1N369A have acceptable performance over the entire band from 10 to 22 kMc.[4]

Radar development work has shown that a mixer can produce substantial noise

[1] "1957 Semiconductor Diode Specifications," *Tele-Tech and Electronic Industries*, **16**, 95, January, 1957, gives properties of 881 manufactured germanium and silicon types. Boon, S. D., "Germanium Diodes," Elsevier, 1956.

[2] Torrey, H. C., and C. A. Whitmer, "Crystal Rectifiers," Vol. 15, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1948. Pound, R. V., "Microwave Mixers," Vol. 16, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1948. Wheeler, H. A., and D. Dettinger, "Measuring the Efficiency of a Superheterodyne Converter by the Input Impedance Circle Diagram," No. 9 in volume "Wheeler Monographs," Wheeler Laboratories, Great Neck, N.Y., 1953. Moxon, L. A., "Recent Advances in Radio Receivers," Cambridge, 1949; see pp. 54–72. Wind, M., and H. Rapaport, "Handbook of Microwave Measurements," Polytechnic Institute of Brooklyn, Brooklyn, N.Y., 1955; see Sec. 15 by B. Small and C. Masucci, "Determination of Crystal Characteristics." Sharpless, W. M., Wafer-type Millimeter Wave Rectifiers, *Bell System Tech. J.*, **35**, 1385, November, 1956. Edwards, C. F., Frequency Conversion by Means of a Nonlinear Admittance, *Bell System Tech. J.*, **35**, 1403, November, 1956.

[3] Miller, P. H., Jr., Noise Spectrum of Crystal Rectifiers, *Proc. IRE*, March, 1947, pp. 252–256. Roberts, Shepard, Some Considerations Governing Noise Measurements on Crystal Mixers, *Proc. IRE*, March, 1947, p. 257.

[4] Feldman, E. J., and S. L. Levy, Broadband Microwave Crystal Developments, *Tele-Tech and Electronic Industries*, **15**, 68, March, 1956.

merely because of the presence of the noise in the applied injection from the local oscillator.[1] A remedy is available in the now conventional use of two mixers so arranged that their desired outputs add but their outputs of noise from the oscillator cancel. The r-f plumbing for this purpose can consist of a magic tee or of a "rat race," either serving to apply the signal and oscillator inputs to one mixer with a given phase relation but to the other mixer with one of the inputs reversed in polarity.[2]

Table 10. Typical Silicon Crystals for U-h-f and Higher-frequency Service

Approx max frequency, Mc	For converter service in superheterodynes				For detector service in "crystal video" receivers	
	Crystal type	Max conversion loss, db	Max noise-power ratio	Impedance as i-f source, ohms	Crystal type	Impedance as video source, ohms
1,000	1N25	8.0	2.5	100-400		
3,000	1N21C	5.5	1.5	200-800	1N32	5,000-20,000
7,000	1N150	6.0	2.0	250-500		
10,000	1N149	5.5	1.5	325-475	1N31	6,000-23,000
16,000	1N78A	7.0	1.5	365-565		
25,000	1N26	8.5	2.5	300-600		
35,000	1N53	8.5	2.5	400-800		
55,000	Experimental wafer type	8.5	1.7-2.2*	250-500		

* Average rather than maximum values.

A slight improvement in crystal-mixer operation, with respect to reduction of conversion loss and of noise, may be obtainable by preventing the loss of energy at the image frequency. Such energy can be present as a result of demodulation between the oscillator second harmonic and the r-f signal and also as a result of the new i-f signal modulating the oscillator fundamental.[3]

Fig. 15. Method of varying inductance slightly in factory for tracking purposes.

UNICONTROL STATION TUNING

58. Tracking Requirements. A superheterodyne receiver having an r-f stage involves two problems to achieve single-dial tuning. One, also present with the t-r-f receiver, is the tracking of the antenna coupling circuit to the r-f interstage circuit in such a way as to accommodate the necessary range of antenna capacitances. The use of a high-inductance primary in the antenna transformer for this purpose is described in Sec. 36.

To secure satisfactory alignment, fixed tuning elements such as coils are usually made with very close tolerances. For this reason coils are sometimes wound in two sections, as shown in Fig. 15. With this construction one or more turns in section *a* can be transferred to section *b* so as to increase slightly the total inductance of the

[1] "1954 Symposium on Fluctuation Phenomena in Microwave Sources," *Trans. IRE*, **ED-1** (4), 1, December, 1954; includes about 24 papers plus summary on pp. 270-273.

[2] Taub, J., and P. J. Giordano, Use of Crystals in Balanced Mixers, *Trans. IRE*, **MTT-2** (2), 26, July, 1954. Strum, P. D., Crystal Checker for Balanced Mixers, *Trans. IRE*, **MTT-2** (2), 10, July, 1954.

[3] Deutsch, S., Optimum Crystal Mixer Operation—The 1N82 Crystal, *Trans. IRE*, **BTR-1** (1), 10, January, 1955. Strum, P. D., Some Aspects of Mixer Crystal Performance, *Proc. IRE*, **41**, 875, July, 1953.

coil. It is evident that this will increase the inductance because in the limit, when all of *a* is transferred to *b*, the coupling between the turns formerly constituting *a* and the remainder of the coil is increased, which increases the total coil inductance. Conversely, if the inductance of the coil is too high, it can be reduced slightly by transferring turns from *b* to *a*. After this factory adjustment a suitable lacquer is applied to hold the turns in place and protect them from humidity. Instead of such close adjustment it is sometimes practical to sort coils in the factory according to their inductance and then use similar ones together; *i.e.*, if the tuning coils of all circuits are slightly high but in corresponding amounts, the single-dial tuning will not be disturbed beyond a slight permissible error in the frequency as indicated on the dial.

The superheterodyne, in addition to antenna tracking, introduces the problem of oscillator tracking, with the requirement that the oscillator circuits differ accurately from the signal circuits by just the amount of the i.f. The oscillator almost always is set at a frequency higher than the signal because this gives a smaller value for the ratio of maximum to minimum oscillator frequency. The smaller value of this ratio is desirable because it permits a much smaller variation of the tuning reactor in the oscillator circuit, which results in a cost saving.[1]

Very large ratios of maximum to minimum frequency in one tuning range in the v-h-f and u-h-f bands, such as 50 to 1,000 Mc, are obtainable by using fixed capacitive electrodes adjacent to suitable movable coils.[2]

59. Tracking with Capacitive Tuning. The simplest method of obtaining oscillator alignment in sets for quantity production is to have a section of the capacitor gang with a smaller number of specially shaped plates so as to produce the required oscillator frequency at all points of the dial. Such a capacitor gang is said to have a "cut-plate" section. In a typical design for an i.f. of 455 kc and covering 540 to 1,630 kc, the signal circuit may consist of a 200-μh coil and a tuning capacitor with a maximum of 420 $\mu\mu$f. The oscillator section consists of an 80-μh coil and a cut-plate section of the tuning gang with a maximum capacitance of 160 $\mu\mu$f. The reduction of oscillator tuning capacitance saves chassis space owing to the shorter length of the capacitor gang and the smaller space required for the emerging rotor of smaller radius. A capacitor gang with a cut-plate section is generally avoided in receivers having more than one tuning band but is frequently found in sets receiving only the standard broadcast band.

Capacitor gangs with identical sections can be used by employing a suitable value of inductance in the oscillator and by employing series and shunt capacitors. A method of computing the proper inductance and proper values for the associated capacitors is given by Roder.[3] This gives a "three-point tracking"; *i.e.*, the alignment is perfect at three points in the band. The choice of inductance, which is a design and factory matter, determines the frequency near the center of the band at which the tracking is perfect. A series capacitor furnishes perfect tracking at a point near the l-f end, and a parallel capacitor near the h-f end. These two capacitors are normally adjustable in the field whenever necessary to restore satisfactory alignment.

A simplified procedure for determining the oscillator inductance and the series and parallel capacitors has been described[4] together with an illustration for coverage of the h-f range from 7 to 15 Mc. Lumping all the minimum circuit capacitance together into a total of 20 $\mu\mu$f and using a capacitor with a maximum of 83 $\mu\mu$f with a signal tuning inductance of 5.6 μh, the oscillator inductance was found to be 5.2 μh, the series capacitor 1,370 $\mu\mu$f, and the parallel capacitor 0.3 $\mu\mu$f.

[1] Dammers, B. G., *et al.*, "Application of the Electronic Valve in Radio Receivers and Amplifiers," Vol. IV of Philips series, 1950, sold in United States and Canada by Elsevier, New York; see pp. 275–307. Sturley, K. R., "Radio Receiver Design," Part I, 2d ed., Wiley, 1953; see pp. 411–434. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 1002–1019.

[2] Lyman, H. T., F. G. Mason, and H. Ross, A Wide Range Tuning System, *IRE Convention Record*, **2**, Part 7, 27, 1954.

[3] Roder, Hans, Oscillator Padding, *Radio Eng.*, **7**, March, 1935. See p. 213 of the fourth edition of this handbook for details.

[4] Marshall, Joseph, Superhet Tracking Formulas, *Electronics*, October, 1946, p. 202.

60. Tracking with Permeability Tuning. In automobile receivers, permeability tuning has largely displaced capacitor tuning. This brings up various questions with regard to the tracking of the oscillator.[1]

There is a limited use of designs in which the variable inductor for the oscillator circuit is identical to the inductors in the signal circuits. In this case, a small series inductor followed by a shunt inductor and by the fixed tuning capacitor can be used to obtain three-point tracking. In a typical example,[2] the tuning inductors had a range from 120 μh at 1,600 kc to 1,000 μh at 550 kc, the signal tuning capacitance being 80 $\mu\mu$f. In the oscillator circuit, the identical inductor was followed by a 26-μh series coil, a 1,127-μh parallel coil, and a 46-$\mu\mu$f fixed tuning capacitor. The i.f. in this design was 455 kc.

Instead of the series and parallel inductors, it is often possible to obtain the effect of the series coil without actually having it present by the expedient of a thicker coil form, and to obtain the effect of the parallel inductor by the use of a tapered winding. This amounts, of course, to employing a different inductance coil in the oscillator circuit. Formulas applicable to this case have been given[3] in the literature.

PUSH-BUTTON TUNING

Tuning by means of push buttons offers freedom from fine tuning. It is an important feature in the design of automobile receivers, and to a less extent with other types of receivers.

61. Mechanical Rotation of Tuning Gang. One available method is to provide an arrangement of mechanical elements by which the user can push a particular button

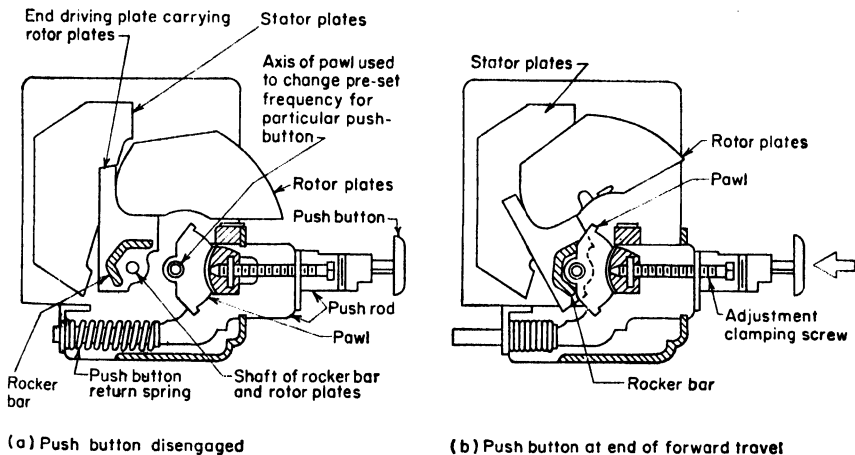



FIG. 16. Rocker-bar type of mechanical push-button tuner.

with sufficient force and through a sufficient distance to rotate the tuning capacitor to the required position. A stop establishing a limit of travel is provided to serve as an indication to the user that the capacitor is at the required position.

A widely used mechanical method of preset tuning has a semicircular pawl or disk which is advanced by the push button, as shown in Fig. 16, the diameter edge of the pawl moving ahead in the direction of the button travel. This pawl encounters a

[1] Fairman, Harry E., Oscillator Tracking Methods in Permeability Tuning, *Tele-Tech*, March, 1947, p. 48. Wessels, P. S., Design of Slug-tuned Superheterodyne Receivers, *Electronics*, **25**, 176, November, 1952.

[2] Simon, A. W., Tracking Permeability Tuned Circuits, *Electronics*, September, 1946, p. 138. See also note by same author in *Electronics*, November, 1947, p. 142.

[3] DeCola, Rinaldo, A Superheterodyne Tracking Solution, *Electronics*, February, 1942, p. 29.

double-arm rocker bar which controls the position of the tuning gang. One end of the pawl encounters the arm on one side of the rocker bar and pushes it ahead until the other side of the rocker bar strikes the other end of the pawl, thus determining the end of button travel and the final position of the rotor of the tuning gang. In this type of mechanism, the preset station for each button can be changed by (1) releasing the adjustment clamping screw, which leaves the pawl free to rotate about its axis; (2) pressing the particular push button and holding it in; (3) tuning in the desired new station with the manual tuning control of the set; (4) tightening the adjustment clamping screw; and (5) releasing the push button. Capacitors with this preset mechanism are designed to go through their entire capacitance range in 90 deg of rotation, and good ratios of maximum to minimum capacitance are obtainable. If a 180-deg capacitor is to be used, a 2:1 stepup gearing is provided.

Another practical method of mechanical push-button tuning has heart-shaped cams on the shaft of the variable capacitor, one such cam for each preset station. Pressure on the push button brings a small roller against the cam and thus rotates the capacitor shaft until the roller ends its travel in the vee of the heart; here the roller is nearest the shaft and can go no farther, and the shaft has been rotated to the desired position. A clamping arrangement allows the cams to be loosened for resetting the stations assigned to one or more of the push buttons.

62. Mechanical Control of Slug Position. Push-button control of permeability tuners can be obtained by mounting the ends of the slugs on a common bar which then moves them into their respective coils. Another arrangement is to use cords to connect the slugs to the tuning dial in such a way that rotation of the dial pulls on one slug, the back end of which is tied to the next slug, the back end of which is tied in turn to the other side of the tuning dial.

63. Multibutton Self-disengaging Switch. Other types of push-button tuning are electrical in character so that the only function of each push button is to perform certain electrical connecting operations. One requirement on a gang switch of this kind is some type of latching mechanism to hold any button pressed and yet so arranged that pressing any button disengages whatever button was previously pressed. Moderate cost is, of course, necessary for these switches to be used in broadcast receivers, and in some types of circuits low capacitance is also required. The self-disengaging feature is obtainable with a latch bar by which the first effect of pushing any button is to disengage the previous button, after which further travel of the button makes the desired connections. The last button operated ordinarily stays more or less depressed to show the user what station is tuned in.

64. Single-button Switch. An extensive use has been made of a single button as the entire preset tuning control available to the user. This has been found acceptable, especially in automobile receivers. Pressing the button causes the set to change from one preset station to the next preset station in a cyclical manner. The receiver in this case includes a solenoid-operated rotary switch which moves around one position for each operation of the single control button. An indicator, such as a numbered or lettered dial behind a window, indicates at all times what preset station is being received.

A more elaborate single-button design can be made, specially devised for autoists traveling in more than one region, in which a spring-operated or motor-operated sweep is employed to bring in the next available station on the air. Each time the button, or bar, is pressed, the tuner progresses to the next station in order of frequency, automatically returning to the other end of the band whenever necessary. There is no presetting of stations, and the receiver can be said to have a "self-readjusting push-button" feature. Further discussion of this type of operation is given in Sec. 153 below.

65. Station Choice by Preset Reactors. A frequently used electrical type of push-button tuning employs preset reactors for the choice of stations. Ordinarily one such reactor is provided for the tuning of the antenna coupling and one for the tuning of the oscillator. In sets having a tuned interstage coupling for continuous tuning, the r-f stage is usually switched out of the circuit for push-button tuning.

The best combination of stability and low cost appears to be the use of preset capacitors for antenna tuning and preset inductors for oscillator tuning. The function of the switch is to select the necessary pair of reactors to tune in the desired station.

66. Motor-operated Tuners. A limited number of sets with motor-operated tuners have been made in which an electrical motor is provided for rotating the tuning capacitor to the required position. The determination of this position, as shown in Fig. 17, is made by providing a disk, which is essentially a two-bar commutator, and several narrow contact brushes, one for each preset station. A contact resting on one-half the disk will cause the motor to turn in one direction, while if it rests on the other half the motor will turn in the opposite direction. With this arrangement, whenever any button is pushed, the motor starts turning in the proper direction to rotate the capacitor shaft directly toward the desired point. As the capacitor rotates, the commutator moves until the insulation between the two halves of the disk is under the contact point, at which time the desired station is tuned in and the motor stops. The setting up of the chosen stations requires only that each contact be positioned around the disk at the necessary angle.

In practice, the contacts are usually arranged in two concentric guide channels. This permits setting two contacts to stations which may be close together in frequency. For setting the contacts, the desired station is tuned manually and the appropriate contact slid around its guide slot until it is on the insulating strip. A fairly high tuning speed can be provided since, if the momentum should cause the tuning to overshoot, the motor will automatically reverse and return to the correct point, where it will stop.

Fig. 17. Motor-operated preset tuner with control disk to start motor in the proper direction to go toward any desired preset station.

67. Frequency Stability. Stability of the various circuits, which is desirable in any receiver, is important for push-button sets. This matter has been the subject of much study and development, and the art is now in a satisfactory state in the sense that almost any desired precision can be obtained at moderate expense. In addition to components less subject to temperature drift,[1] compensating elements have been developed. Within the receiver the problem has been attacked by the use of a wider top i-f selectivity curve and in some elaborate receivers by the provision of a.f.c.

68. Receiver Silencing. Tuners which move the capacitor or inductor gang from one position to another should be provided with a contact for disabling the receiver during the time the change of tuning is in process. Otherwise a disagreeable series of sounds from the various stations passed through in the tuning will be heard. This does not apply to tuners that operate by making a quick change of preset reactors. A

[1] Mistic, G., Temperature Compensating Capacitors, *Elec. Manufacturing*, **57** (2), 83, February, 1956. Sturley, K. R., "Radio Receiver Design," Part I, 2d ed., Wiley, 1953; see pp. 390–404.

ground on the grid of an audio stage can be used to quiet the receiver where necessary during the change of tuning, the arrangement being such that silencing is accomplished in an automatic and economical manner.

69. Remote Control. Some push-button tuning systems lend themselves to convenient employment for remote control, if choice among preset stations, without continuous tuning, is sufficient. The single-button tuner is an outstanding case of this kind since only two wires are required for the tuning. Motor-operated types can also be used for remote control but require a multiconductor cable. Further treatment of remote control is given in Sec. 105.

70. Preset Tuning at Higher Frequencies. Various preset tuners have been devised for aviation, maritime, and military equipment operating at frequencies above the standard broadcast band. In Western Electric equipment, a telephone dial is used to tune quickly to one of as many as 8 or 10 preset frequencies. Another tuner, the Collins "Autotune," has been used extensively in military equipment. This is a motor-operated type which has for each preset frequency a circular disk with a gap into which a detent pawl enters, thus stopping the rotation of the shaft at the required point. It makes a change from one preset tuning to another in 5 sec or less.

I-F AMPLIFIERS

Numerous factors must be considered[1] in the design of the i-f amplifier, particularly in equipment for special services. These factors include the choice of the intermediate frequency, number of stages, gain and selectivity required, whether variable selectivity is to be provided, whether stagger tuning is necessary, etc.

71. Choice of I.F. The chief considerations are (1) better image rejection is obtained with higher values and (2) better gain and selectivity in the i-f amplifier are obtained with low values. With broadcast receivers, the most widely used i.f. is 455 kc; for automobile receivers either 455 or 262 kc is used. For the f-m broadcast band (88 to 108 Mc) the i.f. of 10.7 Mc is generally used. In television receivers the value of the picture carrier in the intermediate stages is either about 46 or about 26 Mc. Radar receivers, with carrier frequencies in the u-h-f and s-h-f ranges, usually employ i-f amplifiers operating at 30 or at 60 Mc. A limited use has been made of i.f.s above 100 Mc.

One of the considerations bearing on the choice of i.f., especially for fixed-frequency operation such as in the point-to-point services, is to avoid values that might cause spurious responses due to high-order frequency conversions in the mixer tube.[2]

72. Typical I-f Amplifier Performance. A single i-f stage is usually sufficient in broadcast receivers, especially if only the standard broadcast band is to be accommodated. Such a stage is usually provided with input and output tuned transformers, each having tuned primary and secondary windings, the coupling being about 1 per cent and the Q about 90, so that the coupling is slightly under optimum, with the advantage that manufacturing tolerances do not result in any of the product having a double-peaked characteristic. The i-f gain of such a stage is ordinarily around 40 db, the band widths about 16 kc at 6 db and 44 at 20 db. The inductance of each winding is generally about 1.0 or 1.1 mh, requiring a capacitance of about 120 $\mu\mu$f.

Sets accommodating the f-m broadcast band usually have 10.7-Mc i.f. and two intermediate stages. The first 10.7-Mc i-f stage usually has a gain of about 29 db and band widths of 240 kc at 6 db and 600 kc at 20 db. The second stage has a gain of about 26 db with greater band widths, representative values being 460 kc at 6 db and 2,100 kc at 20 db. Composite i-f transformers can be used, in which one metal can houses both the 455-kc and the 10.7-Mc elements.

[1] Seal, R. K.-F., Miniature Intermediate-frequency Amplifiers, *Nat. Bur. Standards Circ.* 548, 1954. Tahan, E., Simplified I.F. Amplifier Design, *Electronics*, **25**, 147, September, 1952. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 1020–1071. Bartelink, E. H. B., J. Kahnke, and R. L. Waters, A Flat Response Single-tuned I.F. Amplifier, *Proc. IRE*, **36**, 474, April, 1948. Adams, J. J., Intermediate-frequency Amplifiers for Frequency-modulation Receivers, *Proc. IRE*, **35**, 960, September, 1947.

[2] Charts for assisting in this evaluation have been prepared by R. S. Badessa, Mixer Frequency Charts, *Electronics*, August, 1946, p. 138.

73. Choice of Inductance. To obtain maximum gain in an i-f stage, the L/C ratio should be as large as possible with stability. Too high a value, obtained with small values of C, however, makes the variations of the input capacitance of the following tube produce excessive misalignment of the tuning circuit. For this reason values of C are seldom below 30 or 40 $\mu\mu$f, corresponding at 455 kc to 4 and 3 mh. Battery receivers, which must have high gain with limited plate voltage, use inductances between 2 and 3 mh.

With sets having two i-f stages, and especially with a-m–f-m receivers where high-gain tubes and three i-f stages may be necessary in f-m reception and where it is desired to avoid the switching that would be required to reduce the number of i-f stages in a-m reception, lower gain on the 455-kc i.f. is needed to avoid instability, and inductance values as low as 200 μh are used.

Unequal primary and secondary inductances, but with equal primary and secondary Qs, are recommended for 10.7-Mc operation. In this way only the tube and other distributed capacitance need be used to tune the primary, while 50 $\mu\mu$f or more can be used for the secondary. Under these circumstances the primary circuit will give high gain and the secondary circuit will accommodate the change of as much as 2 $\mu\mu$f in the input capacitance of the following tube over the range of a-v-c action. A coupling slightly under optimum is recommended.

As i.f.s of 10 Mc and higher are considered, the gain which can be obtained is limited by one of three factors. One of these is band width, the difficulty being that if the gain is increased beyond a certain point, the band width will be narrowed; *i.e.*, for a given band width, other conditions being the same, there is a limit to the gain. The second limiting factor is stability, which depends on the grid-plate capacitance. If the gain exceeds a certain amount, sufficient voltage will be fed back to produce regeneration or oscillation. The third factor which may limit the gain is the input loading of the following stage. The losses due to transit time, cathode-lead inductance, and other causes are equivalent to a relatively low resistance shunted from grid to ground across the input of the following stage. This limits the gain that can be obtained.

74. Special I-f Coupling Systems. I-f transformers of various special kinds have been proposed or used in limited numbers.

Single-tuned transformers, with very closely coupled windings one of which is tuned, are less expensive than double-tuned designs and can be used where a limited amount of selectivity is sufficient. A triple-tuned design, having three windings and three capacitors, gives very good performance with a relatively wide flattop and sharp skirt selectivity. This type has been used to a limited extent with push-button receivers, the wide top accommodating some drift.

An interesting i-f coupling of a resonant mechanical type[1] is available, in which the mechanical equivalent of a band-pass filter is obtained by the use of small resonant metal plates or cylinders with suitable mechanical coupling elements. A typical design has a pass band of 8.5 kc, but affords an attenuation of 60 db at 10 kc away, *i.e.*, with a band width of 20 kc. Coupling into and out of the mechanical filter is by means of nickel magnetostrictive elements.

Another special class of i-f transformers is made for variable selectivity, usually having an additional winding which can be switched in or out to vary the coupling and thereby the band width. Further data on variable selectivity are given in Sec. 151.

For very sharp selectivity, such as for telegraph reception, a quartz-crystal filter is often used in the i-f amplifier. A description of this feature is given in Sec. 176.

[1] Hathaway, J. C., and D. F. Babcock, Survey of Mechanical Filters and Their Applications, *Proc. IRE*, **45**, 5, January, 1957. DeWitz, G. H., Consideration of Mechanical and LC Type Filters: Part I, *Trans. IRE*, **CS-4** (2), 54, May, 1956. DeWitz, G. H., and R. I. Scibor-Marchocki, Consideration of Mechanical and LC Type Filters: Part II, Two Criteria for the Comparison of Filters, *Trans. IRE*, **CS-4** (2), 58. George, R. W., Electro-mechanical Filters for 100-Kc Carrier and Sideband Selection, *Proc. IRE*, **44**, 14, January, 1956. Lapin, S. P., Electro-mechanical Filters, *Proc. Natl. Electronics Conf.*, **9**, 353, 1953. Doelz, M. L., and J. C. Hathaway, How to Use Mechanical I. F. Filters, *Electronics*, **26**, 138, March, 1953.

The use of band-pass filters, including M-derived sections,[1] affords sharp skirt selectivity for receivers in point-to-point practice. The over-all band width of one such amplifier measured 12½ kc at 6 db and only 19 kc at 90 db.

75. Proportioning of A-v-c Bias. In intermediate amplifiers of more than one stage, it may be advisable to reduce the amount of a-v-c bias applied to the last stage. This stage has the function of delivering the full amount of signal desired at the detector and should, therefore, not be biased back farther than this function permits. Otherwise distortion will be introduced, especially for high values of modulation.

76. I-f Amplifiers Using Transistors. The use of transistors in 455-kc i-f amplifiers is discussed in Sec. 38. For lower i-f values, especially the 262-kc option for automobile receivers, transistors are readily usable. For higher i.f.s, such as the 1.6 Mc sometimes used in amateur practice and the 10.7-Mc standard value for f-m broadcast receivers, care is needed to choose suitable transistor types and suitable circuit arrangements with regard to coupling systems, neutralization, a.v.c., etc.[2]

77. Wide-band I-f Amplifiers. The i-f amplifier in receivers for television, radar, and pulse-type communication systems must handle a wide band of signal frequency components. This requirement is met by (1) the use of a high value of i.f. to keep the relative band width within reasonable limits, and (2) the use of stagger tuning or of selective feedback to afford a wider band than otherwise obtainable.[3]

As an example of stagger tuning, a radar-type system proposed as an aid to air navigation included a receiver with a 60-Mc i-f amplifier of seven stages tuned individually at 55, 60, or 66 Mc so as to give an over-all band width of 12 Mc and a gain of 85 db, or an average of 12 db per stage.

The use of negative feedback in a wide-band i-f amplifier[4] is illustrated by a 60-Mc design of six stages using 6AK5's grouped into a first triplet consisting of the first three stages followed by a similar second triplet. In each triplet the interstage couplings consist of very closely coupled transformers poled to deliver to the following grid a signal of the same instantaneous sign as on the preceding plate. The negative feedback is then obtained by connecting 33,000-ohm resistors from plate to plate, two such resistors being required in each triplet.

An improved pentode (type 404-A) having less noise and about twice the ratio of transconductance to the total of input and output capacitances in comparison with the 6AK5 is used in the 65-Mc i-f amplifiers in the 4,000-Mc relay system operating between New York and Boston.

Receivers for searching for enemy radar signals included designs[5] with high i.f.s. The highest reported was 300 Mc, the amplifier consisting of ten 6AK5 stages with stagger tuning. A band width of 20 Mc at the peak and 24.5 Mc at 6 db was obtained.

MANUAL AND AUTOMATIC VOLUME CONTROL

78. Manual Control. The function of the volume control is to reproduce the modulation at the desired loudness and preferably to maintain this loudness as the receiver is tuned to different stations of various field strengths or as the field strength of the tuned station varies because of fading. The almost universal general solution to these problems is the use of a.v.c. for securing the proper amount of h-f amplification

[1] Moore, J. B., and H. A. Moore, I-f Selectivity in Receivers for Commercial Radio Services, *RCA Rev.*, January, 1940, p. 319.

[2] Greene, R. D., Characteristics and Some Applications of Fused Junction PNP Germanium Transistors for High Frequency Use, *IRE Convention Record*, **3**, Part 3, 138, 1955.

[3] McWhorter, M. M., and J. M. Pettit, The Design of Stagger-tuned Double-tuned Amplifiers for Arbitrarily Large Bandwidth, *Proc. IRE*, **43**, 923, August, 1955. McLeish, C. W., "Design Procedure for Very Wide Band Radio-frequency Transformers," Report 3317-ERA-272 of National Research Council of Canada, 1954, Ottawa. Barbiere, D., Design Equations for Broad-banding by Stagger Tuning, *Electronics*, **26**, 220, October, 1953. Lebenbaum, M. T., Stagger-tuned I-f Design, *Electronics*, **23**, 114, August, 1950. Moxon, L. A., "Recent Advances in Radio Receivers," Cambridge, 1949; see pp. 84–106. Valley, G. E., Jr., and H. Wallman, "Vacuum Tube Amplifiers," Vol. 18, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1948.

[4] Bode, H. W., "Network Analysis and Feedback Amplifier Design," Van Nostrand, 1945.

[5] Lebenbaum, M. T., Wideband I-f Amplifiers above 100 Mc, *Electronics*, April, 1947, p. 138.

and the use of a potentiometer in the audio amplifier[1] for establishing the output volume. Ideally, this method produces a signal at the detector which has the same strength for all stations and fading conditions, so that after the audio volume control is adjusted, no further attention is required.

The manual volume control in a-v-c sets is almost always located between the diode detector and the first audio stage. Usually the one or two diodes for the detection and a triode as first audio amplifier are in a common envelope, *e.g.*, the 12AV6 and the 12BF6. The most common value of the volume control is ½ megohm, with 1 megohm also frequently found. In best designs, the return circuit from the diodes includes first a tweet filter of about 100 $\mu\mu$f followed by a resistor of 47,000 ohms followed by another 100-$\mu\mu$f capacitor, after which the circuit divides into the a-v-c bus and the volume-control potentiometer.

A capacitor is sometimes provided to prevent the flow of d.c. through the volume control, the aim being to increase the life of the volume control before noise develops. A capacitor is generally used in the lead from the arm of the volume control so as to prevent applying the a-v-c bias to the first audio stage.

A special effect as the volume control is reduced can be obtained with a tone-compensated design[2] in which the apparent strength of the low frequencies is maintained. In the absence of this feature, the low notes disappear and make music sound thin whenever the volume control is reduced. The circuit in Fig. 18 furnishes a path to ground for the higher audio frequencies which is more effective when the volume control is set at the lower positions; the relative strength of the bass notes is thus increased. Such bass-compensated tone controls have been generally found to afford more pleasing reproduction at low volumes and have therefore been widely used in medium- and high-grade receivers.

FIG. 18. Bass-compensated volume control.

Tone compensation for the high notes can also be provided, using the same fixed tap as for low-note compensation. For this purpose, a fixed capacitor is connected in shunt with the upper end of the volume control. This provides a path for the high audio frequencies, delivering them in greater amount at the fixed tap so that they are more effective when the volume is set low.

79. Automatic Volume Control. Practically all radio receivers now manufactured are provided with a.v.c.[3] Another name for this feature is *automatic gain control* (or *a.g.c.*). The usual practice is to employ gradual cutoff tubes for r-f and i-f amplification and a suitable type of converter so that the bias on all these tubes can be made more negative and the gain thus be decreased as the signal strength increases, the bias consisting of d-c voltage developed in a diode detector by the signal itself.

As discussed in Sec. 15, the extent to which an a-v-c system delivers uniform output for various input intensities is indicated by the "a-v-c figure of merit," which is the range in decibels over which the input can vary while producing a change of only 10 db in the output. For home receivers, the a-v-c figure of merit normally lies in the range of 40 to 55 db, being larger for the better receivers.

80. Typical A-v-c Circuits. Figure 19 shows the type of a-v-c circuit now in widest use. It is, of course, a compromise between performance and cost. In it one diode circuit and one load resistor are used to develop both the a-v-c bias and the audio out-

[1] EIA, Standard REC-121-B, "Variable Control Resistors," February, 1953; covers standard values, tapers, taps, tolerances, and mechanical details.

[2] D'Orio, P. A., and R. DeCola, Bass Compensation Design Chart, *Electronics*, October, 1937, pp. 37–38.

[3] Oliver, B. M., Automatic Volume Control as a Feedback Problem, *Proc. IRE*, **36**, 466, April, 1948. Nolle, A. W., Adjustment Speed of Automatic-volume-control Systems, *Proc. IRE*, **36**, 911, July, 1948.

put; *i.e.*, a common detector is used for both a-v-c and audio requirements. In operation, the positive peaks of the i-f signal on the diode plate attract electrons from the cathode. These flow through the secondary winding of the i-f transformer, through the 47,000-ohm filter resistor, and through the ½-megohm volume control, returning to the cathode. When modulation is present, this flow of electrons has both d-c and

FIG. 19. Typical a-v-c circuit in five-tube receiver.

a-f components. The desired amount of a-f voltage is picked off by the tap and applied to the grid of the triode which serves as the first audio amplifier. The rectified d-c voltage across the 500,000-ohm diode load acts through the 2.2-megohm filter resistor in the a-v-c bus to determine the grid potential of the two h-f tubes which are controlled. As the signal strength at the antenna increases, more signal will be rectified by the diode, and a greater negative bias will be applied to the two controlled tubes. This change of bias reduces the gain of these tubes with the result that the output of the receiver changes by a much smaller ratio than the input.

FIG. 20. Essential features of "gas gate" in an a-v-c circuit.

81. Use of Gas Gate with A.V.C. Extensive use is made of a circuit feature called a "gas gate," which offers insurance at negligible expense against a certain type of trouble. Sometimes the converter tube, especially in sets having an r-f stage followed by an untuned coupling, will develop sufficient ionization so that positive ions going to the control grid and thence flowing to ground through the high bias resistor will maintain such a large positive d-c drop in the bias resistor (30 or 40 volts) that the tube will draw an excessive sustained plate current and damage will result. This condition can occur with tubes which are normal as indicated by all the usual tests and circuit applications. The gas gate, as shown in Fig. 20, normally employs the second diode plate and connects it to the a-v-c line on the side of the filter resistor which is toward the grid return of the converter stage. Under these circumstances, any positive potential on the grid causes a flow of current through the low

forward resistance of the gas-gate diode, which prevents the grid from going appreciably positive. Television engineers will recognize the similarity of this to the reinserting diodes used in studio and receiving equipment.

82. Different Selectivity for A-v-c and Audio Detection. Figure 21 shows the use of separate diode plates for the a-v-c and audio detection. This allows the a-v-c diode plate to be supplied with signal from a point in the receiver which is preceded by less selectivity than is used ahead of the audio detector. In fact, the a-v-c diode is supplied from the primary of the last i-f transformer while the audio diode is supplied from the secondary. The advantage of this arrangement is that, when the receiver is tuned through a station, the bias voltage is held about constant so that the station "naturally" becomes louder owing to the additional selectivity in the audio

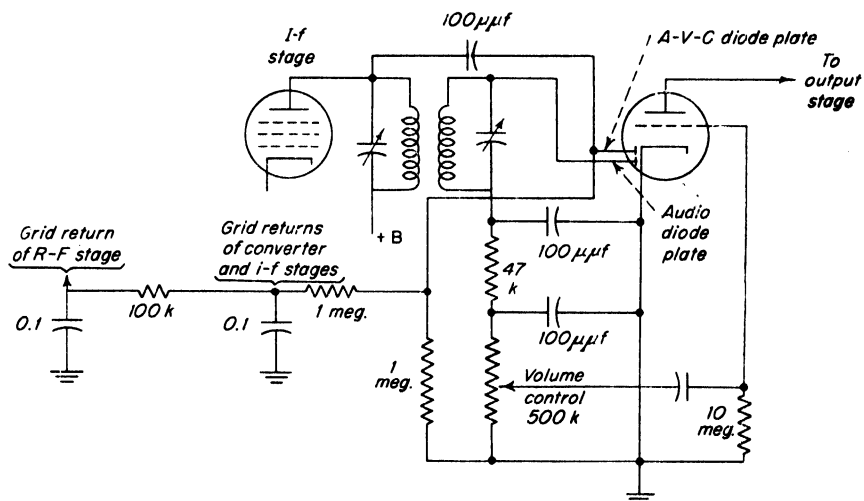



FIG. 21. Use of separate diodes for a-v-c and audio detection.

detector channel and then fades out, permitting easy adjustment of the tuning. Also the large amount of noise heard near each station is reduced. The difference in the signal strength in the two channels should not exceed 20 db. Otherwise the sensitivity in the neighborhood of a strong station may be reduced to the point of preventing the reception of a weak station on the next channel.

In Fig. 21 the a-v-c diode is supplied through a capacitor of 100 μμf and has a load resistor of 1 megohm. No filter capacitor can be connected at this diode because it would detune the primary of the transformer. Instead, filtering is provided by a series resistor of 1 megohm and a 0.1-μf capacitor, after which the control bias is delivered to the grid returns of the converter and i-f stages. Additional filtering is provided for the lead which delivers bias to the r-f stage.

83. I-f Filtering. It is necessary that the i-f signal at the detector be prevented from proceeding on into the audio amplifier and also from going back into earlier stages of the receiver. The presence of i-f signal in the audio amplifier is objectionable because of (1) the likelihood of producing overloading with consequent audio distortion, and (2) the increased probability of feedback of the i-f fundamental or its harmonics into the r-f or i-f stages. Feedback into these earlier stages can occur directly from the detector in various ways, such as due to poor placing of parts, lack of shielding, or need for increased filtering such as for the a-v-c bias line. Feedback to the r-f circuits will produce whistles (or tweets) on stations having frequencies near the har-

monics of the i.f. Feedback of the fundamental into the i-f circuits will produce instability.

To confine the i-f signal the circuit of Fig. 19 includes a filter in the diode plate return consisting of two 100-$\mu\mu$f shunt capacitors and a 47,000-ohm series resistor. This filter greatly reduces the amount of i-f signal getting either on the a-v-c line or on the first audio grid. In Fig. 21 a similar filter is provided in the audio circuit and increased general filtering is included in the a-v-c line.

84. Delayed A.V.C. All the a-v-c circuits just described are subject to the objection that a fairly weak signal will develop some negative bias in the load resistor of the a-v-c diode circuit and thereby reduce the r-f amplification to some extent, all under conditions where the full amplification and the maximum output from the receiver are desired.[1] Various systems have been devised which prevent this effect. In terms of a signal of increasing field strength, these circuits delay the start of a-v-c action until the output has reached a volume as great as desired. The name "delay" is, therefore, with regard to increasing field strength and not with regard to time.

Fig. 22. Various a-v-c characteristic curves.

The difference is shown in Fig. 22 where curve *A* is the characteristic which the receiver will have without any a.v.c. The curve *B* is the characteristic with simple a.v.c. The curves *C* and *D* show the characteristics obtained with delay, more stages being controlled in the case of *D* than with *C*. Curve *E* is the ideal, where the output is uniform above the necessary minimum value of signal strength.

85. Biasing Controlled Tubes in Different Amounts. To minimize a type of distortion frequently encountered due to curvature of the characteristic of amplifying tubes, it is desirable as a general principle to proportion the a-v-c bias on the various tubes so as to have less bias on those operating at higher signal levels.

This is a refinement found in the better receivers. The method consists of a tap on the a-v-c load resistor so that, *e.g.*, in the case of two i-f stages, the last i-f stage receives only half the bias applied to the other tubes of the receiver. The cost consists chiefly of one additional resistor and capacitor for filtering the new a-v-c line.

86. Separate A-v-c Amplifier. For very high-grade performance, such as in communication receivers and in point-to-point practice, a separate a-v-c amplifier should be provided. In one form this consists of a separate i-f amplifier followed by a separate diode, so that greater negative biases are obtained and a flatter a-v-c curve is secured. In the other form the a-v-c bias is obtained in the usual way, and the separate a-v-c amplifier consists of a d-c amplifier which increases the magnitude of the bias voltage before application to the controlled tubes.

The usual a-v-c system is sometimes described as "regressive" to indicate that the control bias is applied to tubes earlier in the circuit than the point at which the diode is located. In distinction to this, the term "progressive" is sometimes used to describe any system in which control bias is applied to one or more tubes located farther along in the circuit than the point where the signal is taken off to operate the a-v-c system. The progressive method can sometimes be advantageously employed where a separate i-f amplifier for a-v-c is provided. A very flat a-v-c curve, *i.e.*, a very high a-v-c figure of merit, is obtained in this way, but care has to be taken to avoid applying too great a bias to the last controlled tube where the signal level may be high. Also, there is a possibility that the a-v-c curve may drop below the hori-

[1] An interesting discussion of remedies for this in point-to-point practice is given by John B. Moore, AGC Noise Considerations in Receiver Design, *Electronics*, May, 1945, p. 116.

zontal, so that the receiver may give less output for greater input signals, if this method is carelessly applied.

Another a-v-c arrangement that will give a very flat a-v-c curve has been used in point-to-point telephone practice[1] and is shown in Fig. 23. In this circuit the a-v-c detector following the a-v-c amplifier provides negative bias for the controlled stages but is also arranged to deliver positive bias to the a-v-c amplifier. The benefit of this positive bias connection can be seen if we suppose that a very large signal is being received and that the output of the set would otherwise be more than desired. The effect of the positive connection is to increase the amplification in the a-v-c amplifier, thus making the a-v-c bias more negative and producing the desired reduction in the amplification of the controlled stages. Conversely, when the signal input is very weak, the positive and negative biases delivered by the a-v-c detector are small so that the gain of the a-v-c amplifier, with a small positive relative bias, is small, with the result that the controlled tubes operate at or near full gain.

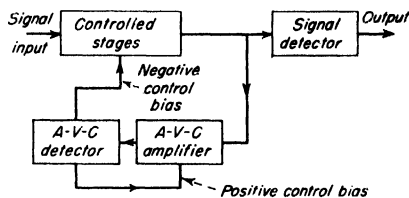
Fig. 23. Use of positive bias in the a-v-c amplifier to obtain very flat a-v-c characteristic.

87. Gain-time Control and Gated A.V.C. In radar receivers the irregularity and the uncertainty of the return echoes make the employment of a.v.c. undesirable, and two other types of operation are used instead. One of these is gain-time control (GTC), in which after each outgoing pulse the receiver is brought up to full sensitivity at a definite rate so that gain increases as a function of time. In this way the powerful echoes from nearby objects are attenuated while echoes arriving later from distant points find the receiver at full sensitivity and are thereby adequately reproduced despite the weak intensity of these echoes. Gain-time control is generally used, for example, to operate a PPI (plan and position indicator) type of oscilloscope, which is the most widely used means of radar presentation.

Some radars are designed to track a given target, such as an airplane, after they are once trained upon it. For this purpose a.v.c. can be used, provided the particular return echo is isolated and used as the control signal. This isolation is accomplished by means of a "gate" which is a time-sampling process by which, at the particular moment when the echo is arriving, the output of the receiver is routed into the a-v-c diode. Here a suitable control bias is developed as in the usual a-v-c circuit. Manual or automatic means are provided so that the timing of the gate is varied to hold the desired target as it varies in its location. Gated a.v.c. therefore operates to control the gain with regard only to a signal arriving at a definite time following the transmission of the outgoing radar pulse. By means of gated a.v.c. the strength of the particular echo can be held constant, independent of wide variations in the intensity of the echo as received on the antenna.

The transmission of radar signals through the usual wave guides makes available an additional means of amplitude control in the form of Faraday-effect rotation of the plane of polarization produced by a d-c magnetic field applied at a ferrite body in the wave guide.[2]

TUNING INDICATORS

A tuning indicator is a device to show visually when a receiver is accurately tuned to a station. Various forms of indicators have been used and two of these are still extensively employed. Those now used are (1) a milliammeter in the plate circuit of

[1] Black, D. M., G. Rodman, and W. T. Wintringham, Ultra-short-wave Receiver for the Cape Charles-Norfolk Multiplex Radiotelephone Circuit, *Proc. IRE*, February, 1945, p. 98.

[2] Vinding, J. P., An Automatic Gain Control System for Microwaves, *Trans. IRE*, **MTT-4**, 244, October, 1956. Loss, M. B., Broadband Applications of Ferrites, *Tele-Tech and Electronic Industries*, **15**, 62, February, 1956.

one or more tubes controlled by a.v.c. and (2) the electron-ray tube, a special form of small cathode-ray tube acting as a voltmeter to indicate when the bias on the a-v-c line reaches its greatest negative value.

88. Meter as Tuning Indicator.[1] The earliest form of tuning indicator, provided in the first a-v-c sets, was a milliammeter showing the current in the controlled tubes. This form is still in general use for communication receivers. Such a meter is often graduated in *S* units to show signal strength according to practice in amateur radio.

89. Electron-ray Indicator Tubes. Various types of small cathode-ray tubes specially constructed for use as tuning indicators have been produced. The most widely used is the 6U5/6G5, which produces a sector-shaped pattern with a total opening of 90 deg when there is zero voltage on the control grid, the angle closing gradually to zero as more negative biases are applied to the control grid. The pattern is completely closed at a potential at −8 volts if the plate supply is 100 volts, at −18.5 volts if the supply is 200 volts, and at −22 volts for a supply of 250 volts. A built-in d-c triode amplifier is included in this electron-ray tube. A-v-c voltage is applied to the grid of the amplifier, which produces an amplified positive voltage on a deflecting electrode internally connected to the triode plate.

90. Advantage of Tuning Indicator. Whether or not to provide a tuning indicator in a given receiver is a question which the engineer may have to consider carefully. For an a-m receiver designed to give high-fidelity reproduction, a moderately wide flat-topped response is desirable, and a tuning indicator should be included so that stations can be accurately tuned to the center of this region, where conditions will be best with regard to freedom from interference. For a high-grade f-m receiver, a tuning indicator should also be provided, especially because, when the set is tuned, the program is generally at a medium or low modulation level, and the general user may have little way of knowing when the tuning is accurately centered. With an f-m set, centering the tuning is more important than with an a-m set because, when loud passages of the program occur in the f-m transmission, the frequency swings over a large part of the 150-kc total deviation. The 6AL7GT tuning indicator has three separate deflector electrodes and allows independent display of the i-f tuning and the discriminator tuning, or other arrangements utilizing the versatility of the tube.

91. The 6AL7GT Electron-ray Tube. This tube[2] with three deflecting electrodes has in addition a control grid which can be used to vary the brightness of the display. This tube does not have a built-in amplifier, but its sensitivity amounts to 1 mm height of display per volt. The display consists of two columns of fluorescence each going upward or downward from a central reference point. The two upper portions are separately controlled by two deflecting electrodes, and the two bottom portions are controlled in step by the third electrode. The grid characteristic has a cutoff at −6 volts, giving rise to the possibility that an indicator-squelch feature can be incorporated so as to darken the display between stations.[3]

92. Other Indicator Tubes. An attractive flared pattern for viewing from the side of the tube is offered by the Philips Type EM 80, which has a heater rated at 6.3 volts and 0.300 amp. Much similar, but with a single sector pattern, is the Lorenz EM 85.[4] Both these types have an over-all length of approximately 2⅝ in. (67 mm).

SQUELCH AND NOISE-LIMITING CIRCUITS

Two methods for combating noise are squelch and noise limiting. Either or both may be used in a given receiver, and both should be considered in special-purpose

[1] Ferguson, R. S., FM Tuning Meter, *Wireless World*, **62**, July, 1956.

[2] Bailey, F. M., An Electron-ray Tuning Indicator for Frequency Modulation, *Proc. IRE*, October, 1947, p. 1158.

[3] For other uses of electron-ray tubes, see L. C. Waller, Applications of Visual Indicator Type Tubes, *RCA Rev.*, January, 1937, pp. 111–125; G. O. Thacker and R. Y. Walker, "Magic Eye, Indicators, *Wireless World*, January, 1947, p. 30.

[4] Malsch, F., Miniaturized Tuning-indicator Tube EM 85, *Elec. Commun.*, **33**, 272, December, 1956.

designs. Extensive use of both features is made in police receivers, and noise limiters are generally included in receivers for the communication service.

93. Squelch Circuits. Various terms are used to designate squelch circuits, such as quiet automatic volume control (or QAVC), muting systems, and codan (*C*arrier-*O*perated *D*evice *A*nti-*N*oise).[1] The purpose of these circuits is to squelch or quiet the receiver when sufficient signal for satisfactory reception is not present. Where the transmission is not continuous but the receivers must always be ready for a signal, as in police service, a squelch circuit performs the important function of removing the large amount of noise which might otherwise be present between transmissions. In receivers for broadcast reception and other services in which tuning through a wide band is necessary, the squelch feature affords the additional benefit of interstation noise suppression.

Considerable care in the design of a squelch circuit is necessary to avoid "ragged-edge" effects when the signal fluctuates rapidly between acceptable and nonacceptable values.

A squelch circuit consists essentially of suitable means for (1) assessing the reception conditions and (2) disabling the receiver as long as satisfactory reception is not possible. The assessing is generally on the basis of conditions in the i-f amplifier, and the disabling is usually accomplished by biasing a grid of the audio amplifier beyond cutoff.

The assessing function of a squelch circuit in an f-m system can utilize the rectified limiter grid current, as representative of signal strength, tending to remove the squelch and let the set operate. At the same time a noise signal of audio frequencies above the transmitted voice range can be rectified and used in the direction of applying the squelch. In police service, when the station carrier comes on for a transmission, the limiter voltage in such a set increases, and the rectified noise voltage decreases; since both effects act to remove the squelch, the receiver becomes operative.[2]

For the application of the large squelch bias a d-c amplifier called the "squelch tube" is generally used. It is ordinarily a triode. The flow of plate current in the triode causes a voltage drop in the plate circuit, and the connections are so arranged that this voltage drop biases the first audio amplifier beyond cutoff, thus squelching the set. Conversely, when the squelch tube is cut off, the first audio stage operates and signals are received.

Since receiving conditions vary, especially with regard to the amount of man-made interference in the locality, an adjustment is provided to control the level required by the squelch to let the set operate. In this way, when interference is slight, the sensitivity and range of the receiver can be increased. Under the best conditions a signal of 140 db (0.1 μv) is sufficient to open the squelch of a police receiver, and the range of adjustment in a typical set is 140 to 128 db (0.1 to 0.4 μv).

Fig. 24. Simple noise limiter of shunt type for headphone use.

94. Noise Limiters. Various arrangements[3] of diodes, multigrid tubes, and other elements are available for limiting the amplitude of noise peaks or disabling the receiver for very short periods while high-amplitude noise is present.

A simple limiter,[4] shown in Fig. 24, utilizes two 1N34 germanium crystals. Each crystal is biased by a dry cell so poled that the crystal becomes conductive whenever the amplitude exceeds the 1.5-volt potential of the cell. The maximum a-c signal

[1] An improved codan circuit is described by H. B. Fisher, Remotely Controlled Receiver for Radio-telephone Systems, *Proc. IRE*, April, 1939, p. 264.

[2] Klehfoth, W. G., A Squelch System Controlled by Signal-to-noise Ratio, *Trans. IRE*, **PGVC-5, 62,** June, 1955.

[3] Toth, Emerick, Noise and Output Limiters, *Electronics*, November, 1946, p. 114; December, 1946, p. 120. Bishop, N., Noise Limiter for Mobile VHF, *Electronics*, **26**, 164, June, 1953.

[4] Grammar, George, Noise Limiting in C. W. Reception, *QST*, May, 1946, p. 13.

at the phones is therefore 0.7 × 1.5, or about 1 volt rms. This limiter is reported to give good results in the reduction of impulse interference, such as automobile ignition, if care is used in the adjustment of the audio and radio volume controls.

Another limiter, widely used in police receivers and in naval shipboard receivers, is discussed in the Toth reference and shown in Fig. 25. This consists of a diode in series with the outgoing audio lead and having such a polarity that it cannot pass the large negative impulses resulting from the rectification of an intense noise train by the detector diode. The operation may be described briefly by saying that a large noise pulse from the i-f transformer, having peaks of both positive and negative polarities, loses the negative peaks in the regular diode detector, and loses the excess positive peaks, now rectified into a negative peak, in the series limiter diode.

The operation of this limiter depends on time constants and particularly on the fact that the cathode of the series diode requires about 200 times longer to change its potential by a given amount than the plate. A sudden negative peak on the plate therefore merely makes the plate greatly negative with respect to its cathode, renders the diode nonconductive, and temporarily open-circuits the audio output connections. By the time the cathode has made a slight start in the negative direction, the noise peak is over and the limiter has served its purpose. For example, the normal potential of the plate may be about −5 volts and the cathode about −7 volts. A noise pulse carrying the plate to −50 volts obviously makes the diode go on open circuit, removing most of the effect of the noise pulse from the output circuit.

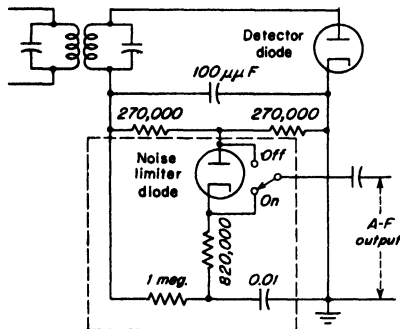



Fig. 25. Series-diode noise limiter.

In summing up a review of several types of noise limiters, Toth concludes that from the standpoint of simplicity, effectiveness, and low distortion, the series-diode type appears best for voice and m-c-w operation. For c-w operation the full-wave a-f shunt type (such as Fig. 24 with the possible substitution of vacuum diodes for the crystals) appears best because of simplicity, effectiveness, and minimum change of gain below the limiting threshold. Both types can be used in c-w service. The benefit of these limiters is greatest where the noise peaks substantially exceed the desired carrier peak values.[1] If this is not the case, only some very elaborate types will afford much relief.

Receivers operating near radars can be blanked by utilizing trigger pulses from the radars. Each of these pulses is properly delayed, and they are then combined and used to disable the heterodyne oscillators of the receivers.[2]

TONE CONTROL

95. Purpose and Types.[3] The low cost and utility of a tone control have made it a common feature of broadcast receivers. It can be defined as any control by which the user can vary the amplitude-frequency characteristic of the audio amplifier of the receiver. Its usual form consists of a simple h-f attenuator. To the average

[1] Cunningham, W. J., S. J. Goffard, and J. C. R. Licklider, The Influence of Amplitude Limiting and Frequency Selectivity upon the Performance of Radio Receivers in Noise, *Proc. IRE*, October, 1947, p. 1021.

[2] Engels, R., An Impulse Interference Blanker for Communications Receivers, *Trans. IRE*, CS-4 (3), 22, October, 1956.

[3] Reed, H., Simplifying Tone Control Circuits, *Radio and Television News*, March, 1956, p. 112. Knight, M. B., Universal Design Curves for Tone-control Circuits, *Radio and Television News*, November, 1951; July, August, September, and October, 1952. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 635-678. Baxandall, P. J., Negative-feedback Tone Control, *Wireless World*, **58**, 402, October, 1952.

listener such a tone control offers the advantages of reducing (1) the worst effects of atmospherics or man-made interference and (2) the high-pitch content of the program if desired.

A survey of 31 models produced during the first two years of the postwar period showed 17 to have tone controls. These consisted of 7 with a series fixed capacitor and variable resistor, 7 with switched arrangements of two to four positions, 2 with independent knobs for bass and treble control, and 1 with multiple push buttons marked for treble, voice, normal, and other types of effects.

The conventional circuit consists of 0.002 to 0.05 μf in series with a variable resistor whose maximum value is usually 500,000 ohms. This circuit is connected from a plate or grid in the audio amplifier to ground or to B+, so as to introduce the desired amount of loss for the higher audio frequencies.

The switched arrangements ordinarily consist of two or three elements, *e.g.*, one capacitor and a fixed resistor so connected that the two may be used in series, or the capacitor only may be put in circuit, or neither may be used.

In high-class receivers, there is an increasing trend toward the provision of separate controls for the bass and treble. The treble control can be similar to the series circuits with a variable resistor as already described. Designs including separate bass and treble controls can be made utilizing negative feedback arrangements. Feedback for tone-control uses may be taken from the high side of the voice coil or from the unbypassed cathode of the output stage, and can be brought back so as to introduce a voltage in series with the volume control at its low-potential end. Series or shunt resistors or capacitors associated with the feedback connection make it possible to obtain desired audio characteristics. A tone-compensated volume control can be included in the circuit at slight extra expense.

AUTOMATIC FREQUENCY CONTROL (A.F.C.)

96. Basic Functions. A radio receiver, already tuned approximately to a given signal, can be automatically brought to almost perfect tuning and be maintained there by means of a.f.c.[1] Drift in the frequency of the superheterodyne oscillator due to warmup can thus be compensated, and also tuning errors due to poor manual adjustment or to inaccuracies in the operation of push-button tuners.

An a-f-c circuit consists of two main parts: the *discriminator* and the *control device.* The discriminator is an error detector and determines the direction and amount of the frequency error to be corrected. The output of the discriminator goes to the control device, which is generally a *reactance tube.* The plate circuit of the reactance tube is made to have a reactive character of controllable amount, so that by connecting it across the oscillator tuned circuit the desired control of the oscillator frequency is obtained. The r-f antenna and interstage circuits are made wide enough not to require correction.

The error found by the discriminator is that existing between the actual signal in the i-f amplifier and the center frequency of this amplifier. The correction made in the oscillator frequency is such that the r-f signal is converted into an i.f. which is much closer to the proper value. Perfect or complete correction with the reactance tube as a control device is not possible because some error is necessary to produce the discriminator output to control the reactance tube. However, this is unimportant in practice because sufficiently complete correction is easily obtainable.

97. Discriminators. The most widely used discriminator[2] has been the coupled-circuit system, also called the phase-variation type, in which the primary voltage is delivered to the center of the secondary, so that the secondary terminals furnish to two diode plates the primary voltage in the same phase on the two plates and the secondary voltage in opposing phases. Both primary and secondary circuits are tuned to the i.f., and this is the frequency which the discriminator uses as its standard of

[1] Foster, D. E., and S. W. Seeley, Automatic Tuning, Simplified Circuits and Design Practice, *Proc. IRE*, March, 1937, p. 289.

[2] Roder, Hans, Theory of the Discriminator Circuit for Automatic Frequency Control, *Proc. IRE*, May, 1938, p. 590.

reference. The essential a-c connections and the action are shown in Fig. 26, where it can be seen that (1) at resonance the two diode plates have equal amounts of signal ($E_{CD} = E_{AD}$); (2) below resonance, plate C gets more signal (E_{CD}) than A; and (3) above resonance, A gets more signal (E_{AD}). This action results from the phase relations normally occurring in two inductively coupled tuned circuits. To obtain a d-c output from the discriminator it is only necessary to add the d-c voltages from the two diodes in a series-opposing connection. The resulting single output then has a polarity indicative of the direction of the error and an amount indicative of the extent of the error.

In f-m receivers with a.f.c., a single discriminator can usually be arranged to produce audio signal, a-f-c output, and a-v-c bias. If only a-f-c output is required, there is

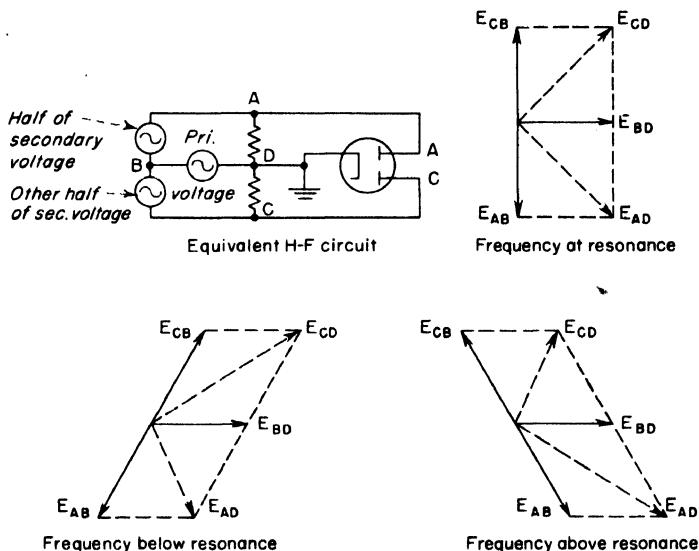



FIG. 26. Action of coupled-circuit discriminator.

greater latitude for proportioning the time constants to secure just the kind of a-f-c action desired. Requirements vary depending on whether the receiver is for fixed-frequency or variable-frequency service, the value of the a-v-c time constant, and whether high-fidelity audio output is required.

For further material on discriminators the reader is referred to the general discussion of f-m detectors in Secs. 140 and 141 and in Chap. 12.

98. Reactance Tubes. The actual shift of oscillator frequency is made by the reactance tube, whose plate circuit is connected across the tuned circuit of the oscillator. The connections can be arranged for the reactance tube to add either inductive or capacitive reactance in parallel with the oscillator tuned circuit. Inductive reactance in parallel raises the oscillator frequency, while capacitive reactance in parallel lowers it.

To obtain a reactive plate impedance, the a-c plate voltage and plate current must be 90 deg out of phase with each other. The plate voltage comes from the oscillator and can be taken as establishing the reference phase. It is necessary that the current be in quadrature with this. Since the current is determined by the grid voltage, the grid and plate voltages must be in quadrature. An RC phase splitter will accomplish this. For example, a small capacitance connected from plate to grid and a small resistor from grid to ground will furnish a circuit from plate to ground which will have an almost pure capacitive reactance. The oscillator voltage will therefore drive a leading current through this circuit. The grid

voltage, being the drop in the resistor, will also lead, and therefore the plate current will lead, all with respect to the alternating plate voltage applied by the oscillator. Since the plate current leads the plate voltage, the plate impedance is capacitive, and the desired control of frequency is obtainable. Ideally C and R in the phase splitter are exceedingly small, and a perfect 90-deg phase shift is obtained. It is desirable that this shift be reasonably near 90 deg so that the plate current contain no substantial component in phase with the voltage. This would represent a loss across the oscillator tuned circuit with the additional objection that it would vary with the amount of frequency correction at the moment.

The magnitude of the capacitance added by the reactance tube in the case just described is approximately $g_m CR$, where g_m is the transconductance of the reactance tube under operating conditions including the bias received from the discriminator. For example, a 6SJ7 reactance tube operating at 1,000 μmhos, with a capacitor of 30 $\mu\mu$f and a resistor of 600 ohms, acts as a capacitor of 18 $\mu\mu$f across the oscillator tuned circuit.

Inductive reactance can be obtained from the reactance tube by exchanging the positions of the two elements, *i.e.*, by using a resistor (with large blocking capacitor in series) from plate to grid and a capacitor from grid to ground. In this case the desired 90-deg relation is most nearly realized with the largest practical values of the resistor and capacitor.

99. A.F.C. in Receivers for the Standard and F-m Broadcast Bands. During the years a limited number of manufactured models for the standard 1-Mc broadcast band have appeared with a.f.c., but ordinarily this feature has been deemed of less value than other improvements of similar cost. This applies to complete receivers and to tuners offered as units for high-fidelity installations. (Automobile receivers of the "signal-seeking" type, as produced since about 1950, include a frequency sweep which carries the tuning up the frequency scale but is subject to automatic stopping whenever a signal is found. These designs do not include a.f.c., and are therefore discussed elsewhere—see Sec. 153.)

An experimental design has been made for the 1-Mc broadcast band using a.f.c. and a sweep feature. In this design the change of stations is made by change of inductance produced by altering the d.c. flowing in an auxiliary winding around the core of suitable saturable material. The permeability of the core is varied by the d.c., and thereby the inductance is changed.[1]

The provision of a.f.c. in an f-m broadcast receiver meets in a very convenient manner for the user the need for accurately tuning to the center of the assigned channel. A tuning indicator is a less convenient method of meeting this need. The reason for this need is that on moderate program levels the signal occupies only a fraction of the channel so that the receiver may unwittingly be mistuned and the listener observe bad quality when high levels occur. Tuners for f-m reception in high-fidelity installations therefore often include either an a-f-c circuit or a tuning indicator.

In an f-m receiver or tuner the control voltage for a.f.c. can be obtained from the f-m detector already in the receiver. It is necessary only to provide the reactance tube and a few resistors and capacitors. An economical solution is to employ a dual triode in which one triode acts as oscillator and the other as reactance tube. The system can be adjusted so as to reduce the tuning error to only $\frac{1}{5}$ to $\frac{1}{10}$ of its previous value.

100. A.F.C. in V-h-f and Higher-frequency Service. The introduction of narrower channels for such services as radio extension of the normal wire telephone system to automobiles has made a.f.c. a necessary feature. Similarly it has been found desirable in receivers for police, fire, and other emergency services. Figure 27 shows the reactance tube and the oscillator-tripler in the General Electric Model 4-ER-4A-1 receiver, designed for fixed-frequency f-m operation on any channel from 152 to 162 Mc and for installation in either cars or headquarters stations. This receiver is a double superheterodyne in which the twenty-seventh harmonic of the crystal is used to reduce the signal to the first i.f. of about 17 Mc. The third harmonic of the crystal is used in the second mixer to produce the second i.f. of 2 Mc. The dis-

[1] Hargens, C. W., III, An All-electronic Signal-seeking Broadcast Receiver, *Trans. IRE*, **BTR-1** (4), 5, October, 1955.

criminator and the general design are arranged for holding the 2-Mc i-f signal centered on the discriminator characteristic.

The 6AK5 oscillator-tripler, shown in Fig. 27, produces the third harmonic for the second mixer. This harmonic is also used after a further multiplication of nine times as the heterodyne frequency for the first mixer. It is slight variation in this third-harmonic frequency of approximately 15 Mc which automatically keeps the receiver in tune and compensates drifting in either the transmitter or the receiver. Without this a-f-c feature, temperature control of the transmitting and receiving crystals would be necessary. The reactance tube in Fig. 27 is half of a 6J6 triode whose plate resistance is increased by operation at a considerable negative grid bias so as not to damp the crystal excessively. As receiver design has proceeded into the u-h-f and s-h-f ranges, a compelling need for a.f.c. has been found. To accommodate the unavoidable drift of the various oscillators, the alternative has been to increase the band width with the

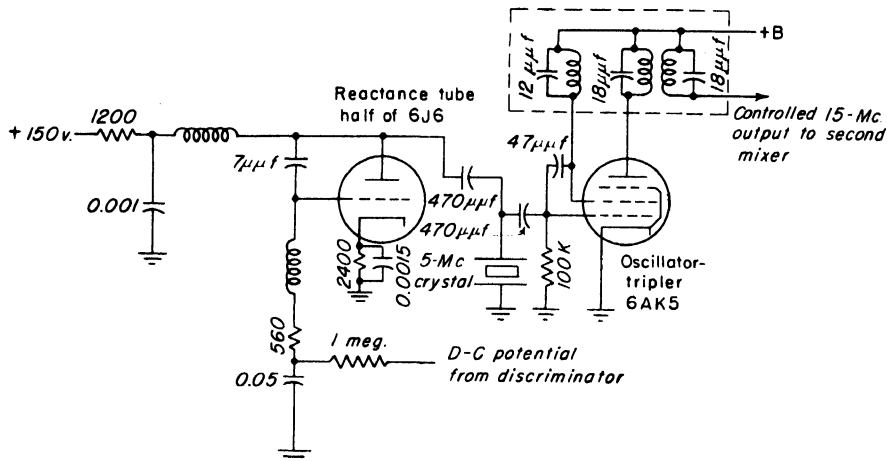



FIG. 27. Use of a.f.c. with crystal oscillator in police receiver.

serious objections of reduced gain and increased noise. Radars and relay systems in these frequency ranges, therefore, make extensive use of a.f.c.[1] One purpose is holding the tuning during warmup, which may be quite important because normal operation may be required very soon after starting the equipment.

Klystrons acting as superhet oscillators can be controlled in frequency by variation of the d-c potential on the reflector electrode. Thus no reactance tube is required, but instead a d-c amplifier to give sufficient values of the voltage to be applied.

The i.f. in receivers for these frequency ranges is generally in the v-h-f band and often is 30 or 60 Mc. A coupled-circuit type of discriminator can be used.

Corrections of as much as 25 Mc may be required and are secured in these services.

101. Mechanical A.F.C. In a mechanical type of a.f.c. the discriminator output polarity determines the direction in which a correcting motor turns, rotating a small air capacitor to the proper point or otherwise making the necessary frequency correction. By gearing the motor down, high accuracy can be obtained. Mechanical a.f.c. has two advantages: (1) the "permanent" nature of the correction, since it does not change in the event of a long fade of the signal—when the signal returns, the tuning is the same and no reestablishment of the correction is required—and (2) perfect correction obtained because the existence of a small error is *not* necessary to maintain the

[1] Jenks, F. A., Simplified Microwave AFC, *Electronics*, November, 1947, p. 120; and December, 1947, p. 132. Miller, Stewart E., Considerations in the Design of Centimeter-wave Radar Receivers, *Proc. IRE*, April, 1947, p. 340. See section on a.f.c. in R. V. Pound, "Microwave Mixers," McGraw-Hill, 1948.

correction. The moving element rotates whenever any appreciable error exists and continues until the error is fully corrected.

The mechanical method of a.f.c. has been used in point-to-point practice[1] and also in radar service.[2]

102. Use of A.F.C. in Double Superhets. The use of a.f.c. in tunable receivers, as for the conventional broadcast band, requires some compromise by the designer to cover from 550 to 1,600 kc (oscillator 1,005 to 2,055 kc). For fixed-frequency service this compromise is not required. Also in double superhets the a-f-c correction can be made on the oscillator supplying the mixer between the two i.f.s, which is a constant-frequency oscillator except for this correction. In these cases very good a-f-c performance can be obtained, especially with regard to the frequency intervals at which the system takes hold and lets go.

SHIELDING AND FILTERING

103. Practical Points. It is common practice to confine the r-f and i-f circuits in metal containers which provide both electromagnetic and electrostatic shielding. In some instances shielded leads are used to provide the connections to the grids or plates of amplifier tubes, but in general the necessity for such shielding is avoided by so locating these leads that they are sufficiently isolated electrically.

In the design of a new receiver the selection of shielding is mostly a matter of general experience and of cut-and-try work in the laboratory. The presence of a shield can around a coil reduces the inductance and increases the losses, so that allowance for these effects may have to be made in the design. For quite low radio frequencies a fair thickness of shield is necessary, the requirement being a thickness of a few times the depth of penetration[3] of induced eddy currents at the particular frequency. The depth of penetration is defined as the depth at which the intensity of the induced current has fallen to 37 per cent of its value at the surface. This figure is $1/e = 1/2.72$, where e is the base of natural logarithms. For aluminum at 100 kc the depth of penetration is 0.27 mm (0.010 in.).

Care must be exercised in locating the power transformer and filter reactor on the receiver chassis, since otherwise the electromagnetic field produced by these units may induce an appreciable hum voltage in the a-f circuits. It is desirable to keep these units separated from the a-f circuits as much as possible, and it is frequently necessary to determine experimentally the best location for these components by connecting them into the circuit with flexible leads and orienting them until a position is established which reduces the hum to the desired minimum.[4]

Resistance-capacitance filters are frequently used in the voltage supply leads for the tube electrodes to prevent coupling between points in the system which differ in signal potential and to provide additional filtering for the voltage fluctuations remaining at the output of the B-supply filter. The d-c drop which can be tolerated in a given circuit is frequently a limiting factor in the use of such filters. Filtering can be improved by dividing the maximum resistance value into two or three series parts and by-passing all junctions to ground. When RC filters are used in circuits in which the average current varies during the operation of the receiver, it is essential that the recovery characteristic of the filter be such that the voltage on the electrode can return to its normal value in approximately 1/10 sec; otherwise noticeable interruptions in the received program will be obtained when sudden changes in the average current occur.

[1] Roetken, A. A., A Single-side-band Receiver for Short-wave Telephone Service, *Proc. IRE*, December, 1938, pp. 1455–1465. Rideout, V. G., Automatic Frequency Control of Microwave Oscillators, *Proc. IRE*, August, 1947, p. 767. van der Wyck, C. F. T., Modern Single-sideband Equipment of the Netherlands Postals Telephone and Telegraph, *Proc. IRE*, **36**, 970, August, 1948.

[2] Stephenson, J. G., Combined Search and Automatic Frequency Control of Mechanically Tuned Oscillators, *Proc. IRE*, **38**, 1314, November, 1950.

[3] Wheeler, H. A., Formulas for Skin Effect, *Proc. IRE*, September, 1942, p. 412. See also Sodaro, J. F., Shielding Nomograph, *Electronics*, **27**, 190, May, 1954.

[4] Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 1196–1201.

This problem is most frequently encountered when *RC* filters are used in the plate or screen circuits of tubes which are controlled by the a-v-c system.

CABINET

104. Purpose and Proportioning. The cabinet must fulfill three requirements:

1. It must house and protect the receiver chassis and loud-speaker mechanism.
2. It must provide sufficient baffle area for the loud-speaker to give the desired l-f response.
3. It must serve as a piece of furniture which will harmonize with the furnishings in the room in which it is to be placed.

Cabinets of wood, plastic, metal, and with leather finish have been found acceptable in the medium and small sizes. Wood and metal are suitable for large cabinets.

Certain proportioning of height and width, in connection with the size, shape, and location of dials and knobs, produces a pleasing appearance, while other proportioning of the various dimensions seems unnatural. A study of these questions has been made.[1]

An important cabinet matter is to avoid the occurrence of acoustic feedback which can arise in receivers capable of producing high power output. To avoid this feedback, the tuning capacitor frequently is flexibly mounted, with respect to the chassis, by means of soft rubber washers. The complete chassis in many receivers is also flexibly mounted in the cabinet.

Acoustic feedback is caused by the loud-speaker vibrations being transmitted to the tuning capacitor or some other circuit element which is thereby caused to vibrate sufficiently to detune the receiver intermittently at an a-f rate. If the proper phase relations exist between the loud-speaker vibrations and the variations in signal intensity which result from the vibration of the capacitor plates, sustained oscillations may be produced.

REMOTE CONTROL AND TELEMETERING

Numerous methods are available for the control of radio receivers at a distance and also for the use of radio receivers for controlling other equipment. A third application of remote control is telemetering in which a distant transmitter controls a local receiver so as to furnish an indication or record of conditions in the neighborhood of the transmitter.

105. Remote Control of Receiver Adjustments. The simplest form of remote control is flexible shafting,[2] which may have a length of 2 or 3 ft for the control of an automobile receiver or a length of 15 or 20 ft for the control of an airplane receiver.

S. S. White flexible shafting for this purpose has considerably less angular deflection under torque than the power type used for automobile speedometers. As an example, a remote-control shaft 0.188 in. in diameter (shaft diameter, not diameter of flexible casing) at a torque of 1 lb-in. has a deflection of 3.3 deg per foot of shafting for either direction of torque. The same size of power shafting under the same conditions has a deflection of 5 deg in the winding direction of the outer layer of the shaft and of 17 deg in the unwinding direction.

The "Flexible Shaft Handbook" gives the torsional breaking strength of the various sizes and types of remote-control shafting. For the 0.188 diam this is about 25 lb-in., so that with a safety factor of 2:1 the maximum working load would be about 12 lb-in. Smaller sizes are offered down to 0.117 in. diam, for which the breaking strength is 4.5 lb-in., corresponding to about 2 lb-in. maximum working torque.

Should the torsional deflection be too great, producing too much backlash, which may well be the case with a length greater than a few feet, gearing should be applied so that the shaft turns at a higher velocity and lower torque than would otherwise be the case. A ratio of 14:1 is common and much higher values are sometimes used.

[1] Van Dyck, Arthur, Dynamic Symmetry in Radio Design, *Proc. IRE*, September, 1932, p. 1481.
[2] "Flexible Shaft Handbook," 3d ed., S. S. White, Industrial Division, 1951.

Another widely used method of remote control is by synchros, which are rotatable machines[1] often called by the trade names Selsyn or Autosyn. One of these is used at each end of the system, with the rotors connected to the same single-phase power-frequency supply line. Each machine has a three-phase stator, the three wires from the one machine being carried the necessary distance and connected to the stator of the machine at the other end. No three-phase power supply is required. If the rotor of either machine is turned, the other rotor follows it closely. Small machines of this kind have been developed by Bendix[2] weighing only 4.25 oz each. A system of four of these machines furnishing control and repeat-back services requires only 15 watts from a 400-cycle power source. In this system, the control machines turn at 10 times the angular velocity of the hand crank and are geared down 600:1 to the remote tuning shaft. In this way the shaft is rotated through 180 deg by 30 complete revolutions of the hand crank. The repeat-back system is geared up 1:1.66 so that at the operating point the 180 deg of shaft rotation appears nicely spread out on a 300-deg dial.

Push-button tuning systems of certain types are adaptable to remote control. This is especially the case with motor-operated systems and with systems in which a rotary switch at the receiver progresses cyclically through various positions for the choice of particular stations. (See Sec. 61 *et seq.*)

In communication services, such as at airports, the level of interference is sometimes so high that it is advantageous to locate the receiving antennas and receivers some miles away. In this case, the telephone line used to deliver the audio signal from each receiver can also deliver d-c voltage to control a reactance tube[3] at the receiver and thus obtain an adjustment of the frequency of the superheterodyne oscillator over a limited range. In c-w telegraph reception, the same method can be used to adjust the beat-frequency oscillator to give the best audio note.

An alternative to the reactance tube, consisting of a coil whose permeability depends on the amount of d.c. in an auxiliary winding, has been described by Polydoroff,[4] who reported an experimental design with a frequency range from about 200 to 350 kc, *i.e.*, 75 kc either side of 275 kc. A similar design for 500 to 3,000 Mc has been reported by Stiber.[5]

106. Use of Receivers for Remote Control of Other Equipment. There has been considerable interest for a number of years in the radio control of model airplanes.[6] The typical receiver circuit on the model plane consists of a self-quenching super-regenerative one-tube receiver using the RK62 thyratron, whose filament is rated at 1.4 volts and 0.05 amp.

Extensive use of radio for the remote control of airplanes was made during the late war.[7] These included guided missiles[8] and also target planes for both aerial and antiaircraft gunnery.

A striking use of radio-controlled planes was made during the Bikini atom-bomb tests by both the Army and Navy, who flew such planes through the cloud above the blast for the collection of samples of the atmosphere to be tested for radioactivity. The Navy planes were modified fighters and the Army planes were B17's. The Navy planes took off from the aircraft carrier Shangri-La, flew through the atomic cloud and then were directed to the

[1] Johnson, T. C., Selsyn Design and Application, *Elec. Engr.*, *Trans. Sec.*, October, 1945, p. 703. Holder, L. F., Principles of Selsyn Equipments and Their Operation, *Gen. Elec. Rev.*, 1930, p. 500.

[2] McNaney, Jos. T., Remote Tuning Unit for Aircraft Radio, *Tele-Tech*, March, 1947, p. 70.

[3] Hannum, E. L., Jr., A Simple Remote Tuning Device for Receivers, *Radio News*, January, 1946, p. 76. Bard, H. B., Jr., Remote Tuning with Reactance Tubes, *Electronics*, August, 1945, p. 100.

[4] Polydoroff, W. J., Incremental Permeability Tuning, *Radio*, October, 1944; also published in *Proc. National Electronics Conf.*, 1944, pp. 146–150.

[5] Stiber, S., Remote-tuning Receiver Has No Moving Parts, *Electronics*, **26**, 186, July, 1953; also in *Proc. Natl. Electronics Conf.*, **8**, 462–468, 1952.

[6] Good, Wm. E., Winning the National Radio Control Meet: Details of a Radio-controlled Model Airplane That Has Made over 100 Successful Flights, *QST*, March, 1940, p. 24.

[7] Radio Controlled Flight, *Radio News*, January, 1947, p. 25. Pilotless Airplane and Missile Control, *Electronics*, April, 1947, p. 152.

[8] Ackerman, S. L., and G. Rappaport, Radio Control Systems for Guided Missiles, *Electronics*, December, 1946, p. 86. Selvidge, Harner, Guided Missiles in World War II, *Proc. Radio Club of America*, October, 1946, p. 3.

island of Roi, a distance of 178 miles, as they were too radioactive to permit their return to the carrier. The remote-control equipment included 10 audio channels with filters and synchro motors for carrying out the various functions. In the same way unmanned boats were sent through the contaminated lagoon by radio to collect samples of the sea water while it was still too radioactive to be handled.

107. Telemetering.[1] A telemetering receiver is a local receiver which brings in signals from a distance representing certain measurements and reproduces these on local indicators or records. Such a receiver is a remote-control device in the sense that it operates the indicator or recorder under the control of the remote transmitter.[2]

An extensive technology exists in the field of detecting, processing, and recording the received telemeter signals. According to the modulation operations at the transmitter, the receiver must perform f-m, a-m, or p-c-m (pulse-code-modulation) detection of the signals conveyed on the carrier and the subcarriers. Also the use of magnetic tape or other form of recording is an important element.[3]

The development of advanced telemetering and remote-control equipment is of special interest in connection with transonic flight, as the unavoidable dangers are thus transferred from a human pilot to a remote-control robot, and the reports of the test are obtained immediately on the ground by telemetering.

MULTIBAND RECEIVERS[4]

A large number of broadcast receivers produced since about 1935 cover, in addition to the normal broadcast band, one or more ranges in the region from 1.6 to 30 Mc.

108. Receivers with H-f Bands. In most receivers with one or more h-f bands, separate coils are employed in the r-f system for each tuning range. A few receivers use single coils with taps for the various bands. When such coils are utilized, the unused portion of the coil is always short-circuited to avoid absorption of signal at higher frequencies. When separate coils are employed, the coils for two or more of the frequency bands are frequently wound on a single form. The coil windings differ considerably with the frequency range. Wire as small as No. 35 AWG is used in the inductances for the tuning range from 540 to 1,600 kc, while wire as large as No. 22 is used in some of the h-f coils. The turns on the h-f coils are usually spaced to minimize coil losses.

These multiband receivers are provided with a gang switch for simultaneously connecting the coils used for each tuning range to the associated tuning-capacitor gang and tubes. On "normal" h-f ranges (as distinct from band-spread ranges) two- or three-gang tuning is used, the circuits being the antenna coupling, the r-f interstage coupling if present, and the oscillator tuning. The band switch connects the proper coil into each circuit for the particular frequency range.

109. Control Knobs and Dials. Multiband receivers are sometimes equipped with tuning mechanisms that permit the user to change the drive ratio between the tuning knob and the variable capacitor from a normal value such as 10:1 to finer control such as 50:1. This higher ratio is helpful in tuning the receiver to an h-f station since the band covered in a single h-f tuning range may be over 10 times that covered from 540 to 1,600 kc.

Special tuning dials are necessary on these receivers since a separate scale is required for each tuning range. In some receivers all the scales are visible, regardless of the tuning range being used, and an indicator or lighting arrangement actuated by the

[1] Nichols, M. H., and L. L. Rauch, "Radio Telemetry," 2d ed., Wiley, 1956.

[2] Redden, M. S., and H. W. Zancanata, Telemetering Receiver Conserves Bandwidth, *Electronics*, **29**, 174, May, 1956.

[3] Warzecha, T. D., Telemetering Demodulator for Wide-band F-M Data, *Electronics*, **29**, 157, July, 1956. Bell, W. S., and C. W. Schultz, Shipboard Telemetering for Terrier Missiles, *Electronics*, **29**, 134, June, 1956. Howard, D. C., UKR-7 Telemetric Data Receiving Set, *Trans. IRE*, **TRC-2** (1), 10, March, 1956. Siderman, J. A., Radiosonde Telemetering and Recording System, *Tele-Tech*, **10**, 40, November, 1951; 54, December, 1951.

[4] Sandeman, E. K., "Radio Engineering," Vol. 2, Chapman & Hall, 1953; see pp. 98–104 and accompanying figure. Maki, G. J., Multiband Tuner Design Chart, *Electronics*, **26**, 200, August, 1953. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 1251–1256, 1285, and 1482.

range switch knob designates the correct scale. On other receivers only the scale corresponding to the tuning range being used is visible. With this arrangement the dial scales are movable with respect to the dial opening, and the range switch is mechanically connected with the dial scales so that, as the tuning range is switched from one frequency band to another, the proper scale is moved into place.

110. Design of Band-spread Circuits. To the amateur radio operator, the use of band spread on the congested amateur bands is of great value. Likewise, the listener to international broadcasts is greatly assisted by the provision of spread bands for these frequencies. In remote parts of the world where there are colonies of American and European citizens dependent on these frequencies for regular broadcast services, the provision of band spread is quite important.

Various arrangements of capacitors and inductors can be used to spread out a relatively narrow band so that it will cover the entire dial of a receiver.[1]

It is often sufficient to provide fixed band-pass circuits for the antenna and interstage r-f couplings, so that only the oscillator circuit is varied as the tuning dial is operated. In this case, Foster and Mountjoy[2] point out that double-tuned antenna coupling gives higher gain and greater uniformity of gain than single-tuned coupling. Typical values for the voltage gain of such a single-tuned circuit covering a complete spread band without adjustment are about 20 db, whereas with a double-tuned circuit values of 24 or 25 db are obtained. In receivers not having an r-f stage, it may therefore be worth while to provide double-tuned antenna couplings on the band-spread positions of the band switch.

An arrangement has been provided in some American models by which two spread bands are accommodated on a single position of the band switch. The lower end of the dial, amounting to about 40 per cent of the total movement, gives the band spreading for one of the bands to be received while the upper 40 per cent gives coverage similarly for the other band. The 20 per cent of the dial between these two parts accommodates in a compressed manner the frequencies lying between these two bands. A special capacitor-plate shape provides these characteristics, the radius changing slowly over the two ends and rapidly in the region between the two bands.

The band-spread ranges on a receiver, to be most useful, must be direct-reading. Since a change of 10 kc at 10 Mc represents only $\frac{1}{10}$ of 1 per cent, it is evident that the stability requirements on the oscillator tuning are very strict. Special attention must therefore be paid to the effects of temperature, age, voltage supply, and other factors on the oscillator frequency to obtain sufficient stability to satisfy the requirements. If temperature-compensating capacitors are used, care is necessary to locate them on the chassis so that the rate of heat transfer to them is correct to match the oscillator drift.

RADIO-PHONOGRAPH COMBINATIONS

The wide interest of the public in phonograph reproduction since the latter years of the 1930's has made this feature one of the most important aspects of receiver design and production. Great numbers of radio-phonograph combinations have been sold, including both table model and console types. In addition, separate electric phonographs have been produced in large numbers, and there has been extensive production of wireless record players as attachments to sets not having phonograph provisions. Numerous technical points[3] have arisen in the design and production of this large

[1] Hughes, D. H., The Design of Band-spread Tuned Circuits for Broadcast Receivers, *J. IEE*, Part III, March, 1946, p. 87, November, 1946, p. 459.

[2] Foster, D. E., and G. Mountjoy, Short-wave Spread-bands in Automobile and Home Receivers, *Proc. IRE*, May, 1942, p. 222.

[3] Jorysz, Alfred, Bibliography of Disc Recording, *Tele-Tech*, June, 1947, p. 73. Robinson, W. H., Lateral Recording, *Communications*, February, 1947, p. 26; April, 1947, p. 38. Goodell, John D., The Reproduction of Disc Recording, *Radio News*, October, 1946, p. 5; January, 1947, p. 3; February, 1947, p. 13. Bachman, W. S., Phonograph Dynamics, *Electronic Inds.*, July, 1945, p. 86. Roys, H. E., Distortion in Phonograph Reproduction, *RCA Rev.*, **14**, 397, September, 1953. Bauer, B. B., Phonograph Reproduction, *IRE Convention Record*, **1**, Part 3, pp. 3-15, 1953. Briggs, G. A., "Sound Reproduction," 3d ed., Wharfdale Wireless Works, Bradford, Yorkshire, England, 1953; see pp. 264-360. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; pp. 700 and 1478.

quantity of equipment, and various special designs have been marketed. The present section summarizes some of the chief engineering data and available methods in this field.

111. Constant-velocity Recording. There are two different fundamental types of recording, constant-velocity and constant-amplitude, a combination of which is generally used.[1]

In constant-velocity recording, if several frequencies are handled in turn by the recording amplifier and are at the same level, each will drive the cutting stylus with the same maximum velocity during each audio cycle. The reproducing stylus will then correspondingly be given the same maximum velocity at the various frequencies when the record is played. Ideally, the magnetic type of cutter produces a constant-velocity recording. This type of recording cannot, however, be used below about 500 cps because it is accompanied by an increase of amplitude which is inversely proportional to the frequency, with the result that the usual spacing of grooves would be inadequate. The crystal type of cutter will produce a constant-velocity recording if it is operated with a suitable network which introduces increasing attenuation at the higher frequencies.

FIG. 28. Arrangement for observing optical pattern of laterally cut phonograph record.

The intensity of recording on a record can be examined as shown in Fig. 28 by reflecting a light from the surface in the proper manner. The pattern seen is defined as the *optical pattern*, often called a "Christmas tree." It has a uniform width for frequencies recorded with the same velocity.[2]

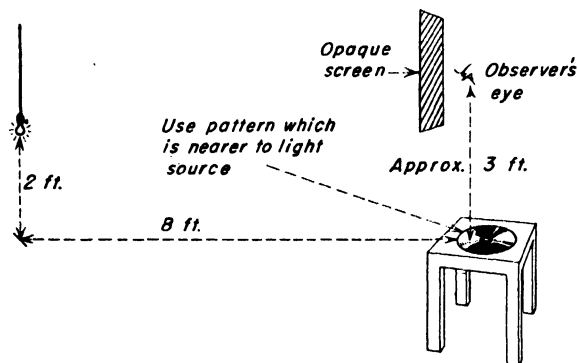
A constant-velocity recording would in the ideal give a faithful reproduction if played with a magnetic pickup. When played with a crystal pickup, it would give faithful reproduction if a suitable network were used to introduce greater attenuation for the lower frequencies.

112. Constant-amplitude Recording. If various frequencies at the same level in a recording amplifier are so transferred to the record that they have the same maximum amplitude on the record, the type of recording and the record itself are characterized as constant-amplitude. A crystal cutter produces a constant-amplitude recording. With a magnetic cutter, a constant-amplitude recording can be obtained by the use of a suitable network with increasing attenuation at lower frequencies.

In constant-amplitude recording, the maximum velocity is proportional to the frequency since the needle must traverse the given amplitude in less and less time as the period is reduced accompanying the increasing frequency. For this reason the velocity with a constant-amplitude recording doubles each time the frequency is doubled. Therefore for each octave change to a higher frequency the constant-amplitude recording has a 6-db increase of velocity. This corresponds to 24 db greater velocity at 8,000 cps than at 500 cps. This can be considered as a substantial preemphasis on the higher frequencies for the purpose of improving the record by minimizing the amount of scratch noise left after deemphasis in the reproduction. The standard recording curve and its corresponding inverse reproducing curve have a general character much nearer constant-amplitude than constant-velocity.

A constant-amplitude recording can be faithfully reproduced with a crystal pickup

[1] See also Chap. 21, Sec. 24 *et seq.*

[2] Bauer, B. B., An Improved Optical Method for Calibrating Test Records, *Trans. IRE*, **AU-3**, 137, September–October, 1955.

having little or no equalization. It can be faithfully reproduced by a magnetic pickup with the use of a proper network with increasing attenuation at higher frequencies.

113. Tracking Error. The reproducing stylus is not operated vertically against the record but rests back at an angle. It is desirable that this angle be directly above the tangent to the groove, so that the top end of the needle is at the same distance from the center of the record as is the bottom end. Departure from this relation is called "tracking error"[1] and is usually kept to moderate proportions by placing the pickup cartridge at an angle on the end of the tone arm, or obtaining the equivalent result by a curved shape of the entire tone arm.

114. Adopted Phonograph Standards. The various standardizing groups (Audio Engineering Society, Record Industry Association of America, EIA, NARTB, etc.)

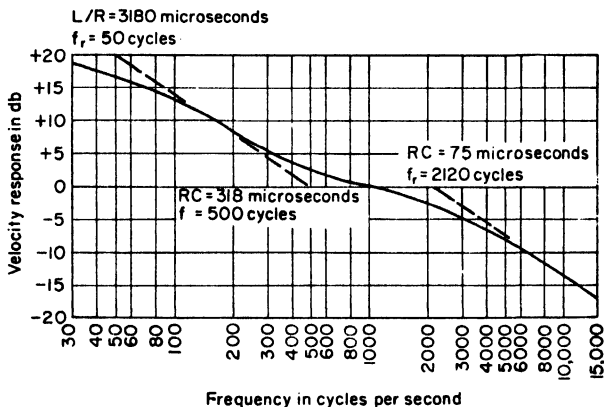



FIG. 29. Standard playback characteristic for phonograph records, showing deemphasis for noise reduction.

have adopted standard dimensions for home disk records,[2] a standard recording and playback characteristic,[3] and other standards in the recording field.

The standard value of the nominal 78-rpm speed is 78.26; this can be conveniently checked by observing a 92-bar pattern illuminated with 60-cycle flashes from a neon lamp.

The standard recording and playback curve provides substantial preemphasis for the highs in recording. Conversely the highs are deemphasized in reproduction, as shown in Fig. 29. The curve is characterized by three time constants associated, respectively, with the low-, middle-, and high-frequency regions of the audio spectrum. The dashed tangents in the drawing have a slope of 6 db per octave, and the curve at each stated value of frequency f_r departs by 3 db from the tangent line.

One of the defects to which turntable mechanisms have been subject in the past is flutter (or "wow"), which is the effect of fluctuation in the speed of rotation.[4] NARTB standards specify that the maximum instantaneous deviation from the mean speed when making a recording for broadcast-station use shall not exceed ±0.1 per cent of the mean speed.

115. Usual Practice in Home Playing of Disks. The usual practice in home reproduction is characterized by speeds of 33⅓,[5] 45, and 78 rpm, record sizes of 12, 10,

[1] Bauer, B. B., Tracking Angle in Phonograph Pickups, *Electronics*, March, 1945, p. 110.

[2] EIA Standard REC-103, "Dimensional Characteristics of Phonograph Records for Home Use," June, 1954.

[3] EIA Standard REC-146, "Lateral Disc Recording Characterstic," September, 1955.

[4] Furst, Ulrich R., Periodic Variations of Pitch in Sound Reproduction by Phonographs, *Proc. IRE*, November, 1946, pp. 887-895.

[5] Goldmark, P. C., R. Snepvangers, and W. S. Bachman, The Columbia Long-playing Microgroove Recording System, *Proc. IRE*, **37**, 923, August, 1949.

and 7 in. diameter, pickup of crystal, magnetic, or ceramic type, and the use of either a simple turntable or a record changer.

For good reproduction the reproducing stylus must be larger than the recording stylus; otherwise the reproducing stylus will rattle about in the groove. It is interesting to note that the contact areas involved are so small that the pressure is very high, probably of the order of several tons per square inch. Under this pressure the groove walls yield elastically when the record is played, but it is found that if the reproducing stylus is well polished and properly proportioned and the stylus force is held to 1 oz or less, the deformation is not permanent and the groove is not damaged even by a large number of playings.

For 78-rpm records the radius at the bottom of the groove is approximately 0.0015 in., and the radius of the needle 0.0025 to 0.0032 in. For 33⅓-rpm records the corresponding groove and needle radii are 0.0005 and 0.001 in.

The popularity of the phonograph has been enhanced by the introduction of low-priced record changers and also by the economy and performance of the rim type of drive in which a large rubber-tired pinion under the turntable furnishes motive power by driving against the turned-down edge of the turntable.

The usual phonograph combination introduces the audio signal from the pickup into the first audio stage in place of the output from the diode detector. The first audio stage is followed by the output stage, the two furnishing a total gain of about 55 db when the volume control is set at full gain. The volume control is usually a potentiometer at the input of the first audio stage and operates for either radio or phonograph use. Separate phonographs, which have been widely sold, include two audio stages and a rectifier and may also have additional tubes to provide a push-pull output stage and to use a tube inverter in place of the push-pull input transformer. In the case of a magnetic pickup, the voltage is low and a preamplifier is provided which for the present discussion is considered part of the pickup.

A limited use is being made of 16⅔ rpm in records for the home in the interest of longer playing time, and "four-speed" players accommodate these records. The usual records for the blind, sponsored by the Library of Congress, operate at 33⅓ rpm. Both these types of records operate with a needle of 0.001 in. radius. (Records for 16⅔-rpm home use differ from 16⅔-rpm automobile records, which are described in Sec. 117.)

116. Special Types of Home Reproducers. Various special disk-playing methods have been devised, such as the modulation of a light beam by a mirror attached to the reproducing needle. Another special arrangement[1] produces a.m. of a 2.5-Mc oscillator by movement of a vane attached to the needle. This design has low mechanical impedance and requires a stylus force of only ½ oz (14 g), so that it produces very little wear of the records. Other designs in which the needle produces f.m. of an oscillator have been described.[2]

117. Phonograph for Automobiles. A phonograph for installation in automobiles has been introduced.[3] By the use of a needle radius of ¼ mil (0.00025 in.), it is possible to space the grooves closer (about 550 per in.) and thus with a 16⅔-rpm speed and a 7-in. record diameter obtain 45 min of music or 1 hr of speech per side. A stylus force of 2½ g is used. The signal and noise are each about 6 db below the values for a conventional 33⅓-rpm record, so that the signal-to-noise ratio is approximately the same.

Vibration and road shocks are controlled by (1) a rubber shock-mount system, (2) a uniform weight distribution, and (3) counterbalancing the tone arm about both axes. In this way the difficulties of wow and change of stylus force are obviated.

[1] Kalmus, Henry P., Pickup with Low Mechanical Impedance, *Electronics*, January, 1946, p. 140; see also the same author's Improved Modulated-oscillator Pickup, *Electronics*, July, 1946, p. 182.

[2] Miessner, Benjamin F., Frequency Modulation Phonograph Pickups, *Trans. IRE*, **AU-2**, 121, July–August, 1954.

[3] Goldmark, P. C., A Phonograph System for the Automobile, *IRE Convention Record*, **4**, Part 7, pp. 159–165, 1956.

Constant turntable velocity is obtained by an induction motor driven from a 60-cycle vibrator.

118. Wireless Phonograph Attachments. One of the standard forms of phonograph equipment is a wireless attachment or phonograph oscillator which produces an a-m wave in the broadcast band for reception by the ordinary broadcast receiver. Many of these designs utilize the l-f end of the broadcast band, being adjustable, for example, from 530 to 730 kc. Others are adjustable similarly at the upper end of the broadcast band, *e.g.*, anywhere from 1,500 to 1,800 kc. These equipments usually consist of a one-tube modulator and a one-tube rectifier. The only external connection is to a 110-volt outlet. This must ordinarily be a.c. owing to the motor power requirement. At the time of installation the frequency of the oscillator is set at an appropriate point well removed from any local broadcast stations.

These equipments are permitted by the FCC to operate without a transmitting license on the condition that the field strength must not exceed 15 μv per m at a distance equal to the wavelength divided by 2π.

119. Frequency Records for Testing. Various firms have made frequency records which are of much value in experimental development, as well as in high-grade maintenance work. Some of these are listed in Table 11.

Table 11. Phonograph Test Records

Company	33⅓ rpm	45 rpm	Type of characteristic
Columbia	No. TL-1	None	Approx RIAA*
RCA Victor	No. 12-5-49	No. 12-5-51	Accurate RIAA*

* RIAA = Record Industry Association of America.

120. Dynamic Noise Suppression. A circuit known as a "dynamic noise suppressor" has been introduced by Scott to improve the reproduction of noisy phonograph records[1] on the following basis. When music is at a low level, the lowest and highest pitches are inaudible. There is therefore no harm done to the quality of reproduction if the system introduces high attenuation for these components. Under these circumstances noise present in the same frequency regions is removed, and the reproduction may sound quieter. This noise suppressor is, therefore, a device which controls the frequency band in the reproduction and varies it automatically in accordance with the volume and frequency characteristics of the music. The frequency band is widened during loud passages and contracted during weak passages, the action being independent at the bass and treble ends of the audio range.

The noise to be reduced consists of scratch in the treble region and motor rumble in the bass region. These noise levels, except for a small portion of the time, may be more intense than the musical components existing in the same frequency regions. It is possible to control each end of the spectrum by means of signals present in a region nearer to the center of the audio range. Control in this way resides effectively in the signal rather than in the noise. With music this is possible, since high harmonics exist only when there are fundamentals of somewhat lower frequency which can be used to operate the control. Similarly at the bass end, whenever deep fundamentals are present they are accompanied by harmonics of appreciable amplitude which can be used to control the cutoff.

A model of this suppressor designed for use in broadcast stations has 13 tubes. It provides 20 db of scratch reduction and about 15 db of rumble reduction. Models suitable for home use add two or three tubes to the normal receiver complement.

[1] Scott, H. H., Dynamic Noise Suppressor, *Electronics*, December, 1947, p. 96. Purinton, E. S., Dynamic Amplifiers for Phonograph Reproduction, *Trans. IRE*, **AU-2**, 80, May–June, 1954.

MAGNETIC RECORDING[1]

121. Home Recorders. There is an increasing interest in the home recording of favorite radio programs as well as the voices of children, members of the family, etc., and in the reproduction of musical pieces from purchased prepared magnetic tapes. Recorder-reproducers for such home use are offered in many models.[2] These utilize magnetic tape of ¼-in. width. The most frequently used speeds are 7½ and 3¾ in. per sec, with some use of 1⅞ and 15 in. per sec and other speeds. The total film thickness may typically be approximately 0.002 in.

122. Ultrasonic Bias. The high quality of sound reproduction afforded by magnetic tape is due in part to the use of an ultrasonic "bias" of 30 to 70 kc which is applied to the tape along with the audio signal. This type of operation substantially improves the linearity and therefore reduces distortion.[3]

VOLUME EXPANSION AND COMPRESSION

It is possible in an amplifier to vary the gain in such a way that the greater the input, the greater the gain. Such an action is called "volume expansion" because the range of volume applied to the amplifier is increased or expanded. Conversely, the gain may be varied in the opposite direction to produce volume compression.[4]

123. Use in Point-to-point Radiotelephony. Volume compression and expansion are a well-established method of improving the signal-to-noise ratio in point-to-point radiotelephone practice. In this field the term "compandor" is used to designate the compressor and expander equipment at the two ends of the radio link. The compressing and expanding operations are performed at audio frequency and require the use of suitable time constants and other refinements.[5]

124. Compression in Broadcasting and in Phonograph Recording. In broadcast practice symphonic music is compressed by strengthening the weak passages with a manual volume control and reducing the loud passages by either manual or automatic means. Such compression is considered necessary to keep up the average percentage of modulation of the station to deliver a usable signal in the outer parts of the service area, and at the same time avoid overmodulation on the strong passages. In recording phonograph records, compression is similarly used to make the music conform to the limitations of the medium. Both the radio and record reproduction are, of course, normally carried out without any compensating expansion.

There is a question as to whether the listener under typical conditions desires the reproduction of a classical program at the full volume range. To have such a true rendition of the music, the listener must be in a quiet place so that he can hear the pianissimo passages without difficulty, and he must also be in such a place that the fortissimo passages will not be objectionable to other persons in the vicinity.

In some situations, *e.g.*, in receivers operating in automobiles (or receivers for speech in locomotive cabs), it may be desirable to introduce compression rather than expan-

[1] Selsted, W. T., and R. H. Snyder, Magnetic Recording—A Report on the State of the Art, *Trans. IRE*, **AU-2**, 137, September-October, 1954. Camras, M., Magnetic Recording, *IRE Convention Record*, **1**, Part 3, 16, 1953. Hull, J. B., A Simplified Method for the Performance Measurement of Magnetic Tape Recorders, *IRE Convention Record*, **4**, Part 7, 75, 1956. Comerci, F. A., S. Wilpon, and R. Schwartz, Navy Standardization of ¼-inch Magnetic Tape and Recorder-Reproducers, *Trans. IRE*, **AU-2**, 146, September-October, 1954. Chap. 21, Sec. 39, of this handbook.

[2] LeBel, C. J., "How to Make Good Tape Recordings," Audio Devices, New York, 1956. Rettinger, M., "Practical Electroacoustics," Chemical Publishing, 1955; see pp. 221–258. Briggs, G. A., "Sound Reproduction," 3d ed., Wharfdale Wireless Works, Bradford, Yorkshire, England, 1953; see pp. 231–263.

[3] Holmes, L. C., and Donald L. Clark, Supersonic Bias for Magnetic Recording, *Electronics*, July, 1945, p. 126.

[4] Purinton, E. S., Dynamic Amplifiers for Phonographic Reproduction, *Trans. IRE*, **AU-2**, 80, May–June, 1954. See also Chap. 15 of this handbook.

[5] Carter, C. W., Jr., A. C. Dickieson and D. Mitchell, Application of Compandors to Telephone Circuits, *Elec. Eng., Trans. Suppl.*, December, 1946, p. 1079. Harp, M. C., *et al.*, Application of Compandors to FM Radio Systems with Frequency Multiplexing, *Trans. IRE*, **MTT-2** (2), 36, April, 1954. Boxall, F. S., and R. S. Caruthers, A Miniature Compandor for General Use in Wire and Radio Communication Systems, *Communication and Electronics* (AIEE), No. 10, p. 804, January, 1954.

sion so that the volume may be held near a level sufficient to be heard easily over the noise and yet not rise to objectionable loudness during high-modulation periods.

125. General Methods. Equipment with an expansion feature has been offered to the public in a few models during past years. These have been intended especially for phonograph reproduction, where the operation is facilitated by the fact that the compression is more systematic than with broadcast radio programs.

One de luxe method of obtaining compression and expansion in a system is to have a separate pilot control channel, which transmits a signal at all times representing the amount of compression at the transmitter and controlling the compensating expansion at the receiver.[1]

In the absence of a pilot control channel, expanders for broadcast and phonograph use must have different time constants for their operate and release actions. In particular, the time constant for taking hold must be quite short, such as 0.001 sec. The release time must be much greater, such as ½ to 1 sec. The amount of expansion can reasonably be about 10 db at the loudest passages.[2]

126. Expansion by Tube Characteristics. One method of obtaining expansion is by control of the grid bias of an amplifying tube.[3] Such control is usually made progressive, or forward acting; *i.e.*, signal is rectified and the resulting bias used to vary the gain of a tube farther along in the system. Either diode tubes or germanium crystals, such as the 1N34, can be used for the development of bias for expansion.

An alternative method of using a tube for expansion is to connect the plate circuit of a triode so that its resistance is part of a gain-control potentiometer in the amplifier. The 6C5G triode has been used in series with a fixed 50,000-ohm resistor for this type of operation.

Two other available methods of using tubes for volume expansion are (1) connecting the plate circuit of a controlled tube in parallel with the plate circuit of the regular tube, the control arrangements being such that at low levels the shunt loss is greater than at high levels, and (2) employing a pentode or heptode with the audio signal on one grid and the control bias on another grid.

127. Expansion by Nonlinear Resistance. The wide variation of resistance of lamp filaments with current or voltage has been used to furnish volume expansion. However, thermistors[4] (resistors whose value varies with temperature) are more useful for this purpose. In a typical case the resistance is 50,000 ohms at a very low electrical level but declines to only 800 ohms when dissipating 100 mw. These thermistors can be used in shunt across the circuit to produce compression or in series in the circuit to produce expansion. By arrangement in a bridge manner, d.c. can be passed through them to produce further control of the amount of expansion or compression, the d.c. being developed from the signal itself.[5]

A striking application and extension of compression and expansion technique was demonstrated in 1941 by Harvey Fletcher and others of the Bell Telephone Laboratories.[6] This included the feature of manual control of the expansion by a musical conductor, who thereby was effectively given an orchestra of a thousand pieces available at the turn of a knob. This possibility of "enhancement" of the original volume range affords very impressive musical effects. A three-dimensional feature was also shown by the use of three channels with speakers at the two sides and center of the stage. The volume range

[1] In this connection see Haynes, H. E., New Principle for Electronic Volume Compression, *J. Soc. Motion Picture Television Engrs.*, **58**, 137, February, 1952.

[2] Topham, J. W., A Volume Expander for Hi-Fi, *Radio and Television News*, November, 1956, p. 54. Blecher, F. H., Automatic Gain Control of Junction Transistors, *Proc. National Electronics Conf.*, **9**, 731, 1953. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; pp. 679–700 and 1478.

[3] Butz, A. Nelson, Jr., Surgeless Volume Expander, *Electronics*, September, 1946, p. 140. Ehrlich, R. W., Volume Expander Design, *Electronics*, December, 1945, p. 124.

[4] Batcher, Ralph R., Thermistors in Electronic Circuits, *Electronic Inds.*, January, 1945, p. 76. Becker, J. A., C. B. Green, and G. L. Pearson, Properties and Uses of Thermistors—Thermally Sensitive Resistors, *Bell System Tech. J.*, January, 1947, p. 170.

[5] Bennett, W. R., and S. Doba, Vario-Losser Circuits, *Elec. Eng., Trans. Sec.*, January, 1941, p. 17.

[6] The equipment and methods are described in seven papers in *J. Soc. Motion Picture Engrs.*, October, 1941, p. 331.

of the "stereophonic" film was 50 db, which was increased to 80 db by the use of 30-db normal compression and expansion. The enhancement was obtained by further expansion when desired. This system included three pilot channels, one for the control of the volume of each of the loud-speakers on the stage, the control frequencies being 1,260, 3,780, and 6,300 cps. The single film included three sound tracks for the three loud-speakers and a fourth track on which the three pilot frequencies were recorded. In the reproduction the three pilot frequencies were separated from each other by audio filters and each then used to control the volume of its assigned channel. Optical-photographic recording was used for the four tracks on the film.

In this demonstration the thermistor type of element was used for both the compression in making the film and the expansion in reproduction.

A-C–D-C AND THREE-WAY RECEIVERS

The a-c–d-c type of power supply is popular because of (1) the economy of not requiring a power transformer and (2) the adaptability to either a-c or d-c power supplies.

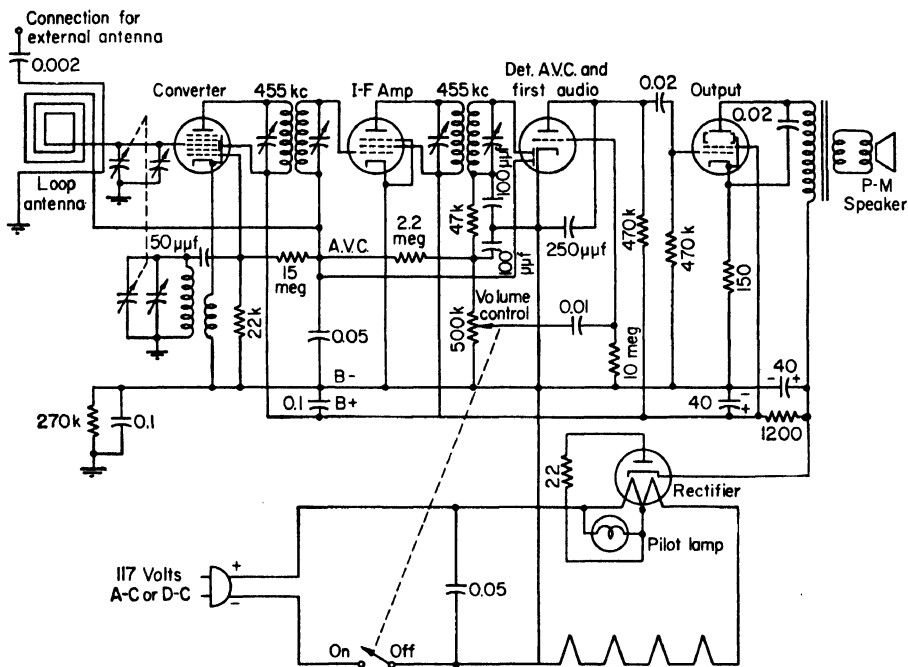



FIG. 30. Typical a-c–d-c receiver circuit.

128. Essential A-c–D-c Features. In the a-c–d-c power system the heaters of the tubes are operated in a series circuit connected across the line, and the plate circuit is operated as an additional load circuit with a half-wave rectifier in series. Figure 30 gives a typical circuit of an a-c–d-c receiver, and it may be noted that the power plug is marked with + and − indications. This polarity must be observed when the set is operated from a d-c outlet. The set is inoperative when plugged into a d-c outlet with the opposite polarity because the tubes have no plate supply. When the set is operated on a.c., heater current flows throughout the a-c cycle, but plate current flows only when the a-c wave has the proper polarity and is near its peak value. The 40-μf capacitor connected from the cathode of the 35Z5GT rectifier to B− maintains most of its charge during the power-frequency cycle so that it is only at positive peaks of the

a-c wave that the rectifier passes current and this capacitor assumes its maximum charge. The action is similar to peak detection.

On account of this pulselike flow of plate current from the a-c line, an a-c–d-c receiver may not operate satisfactorily when an external series resistor is used to adapt a 110-volt set to a higher a-c socket voltage, such as 220 or 250 volts. Separate resistors for the A and B supplies, or other special arrangements, may be necessary.[1]

Selenium rectifiers are now available as a substitute for the tube type and offer the advantages of (1) greater compactness, (2) less production of heat, (3) longer life, and (4) in the usual three-way receiver (operating on a-c, d-c, or battery power) immediate starting of reception after the receiver is turned on.

Capacitor-input filters are always used in a-c–d-c receivers. If the first capacitor in a set with a tube rectifier materially exceeds 40 μf, the set will be subject to failure whenever the user turns it off and then quickly turns it on again. Under these conditions the filter capacitor will quickly discharge, the rectifier cathode will remain hot for the short time involved, and the result will be that, when the set is turned on again, the tube current limit of 500 ma may be greatly exceeded by the rush of current into the capacitor, ruining the rectifier tube and possibly other parts of the receiver. The design in Fig. 30 includes a 22-ohm resistor in series with the rectifier plate as a protection against this effect. This resistor often has a fuselike type of construction so that, if overloaded owing to a short circuit anywhere on the plate-supply system, it will burn out without damaging the rectifier tube. A study of these matters made by Tung-Sol led to the recommendation of the resistance values shown in Table 12 for protection of the rectifier tube.

Servicemen have been advised to install a 50-ohm resistor in all a-c–d-c sets not already equipped for the protection of the power rectifier. This gives sufficient protection unless a very large first filter capacitor is used.

Higher plate voltage in an a-c–d-c set when operating on a-c supply can be obtained by the use of a voltage-doubler rectifier circuit.

Since a-c–d-c sets are usually housed in small cabinets of limited acoustic properties, various suggestions have been made for improving the audio quality and in particular for improving the apparent amount of bass in the sound. One arrangement advanced by F. H. Shepard, Jr., is to introduce a small amount of distortion at all volume levels in the audio amplifier, there being about 5 per cent of third harmonic always present and a small amount of fifth harmonic. In this way the output-input characteristic is changed from a straight line into a slightly S-shaped curve. The resulting distortion is stated to give the listener the impression that the set has fair output at quite low audio frequencies. This feature was demonstrated to the Radio Club of America in September, 1941, and has been described briefly in the literature.[2]

Table 12. Series Resistance Values for Protection of Rectifier Tube

Capacitance of First Filter Capacitor, μf	Series Resistance, Ohms
60	28
80	45
100	65
120	85

129. Three-way Receivers. The addition of circuits for battery operation to an a-c–d-c design produces a convenient form of receiver, usually called a *three-way* type. Various arrangements of switches are suitable for adapting the receiver to either the socket-power outlet or to the batteries carried internally. In one method battery operation is obtained by "plugging in" the normal power cord at a socket on the receiver, this action serving to operate mechanically a switch which connects the A and B batteries into the circuit—*i.e.*, for battery operation no electrical contacts

[1] This condition has been encountered in England and is discussed by F. Langford-Smith in "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952, pp. 1264–1266.

[2] Improved Bass for Small Radios, *Electronics*, July, 1945, p. 224.

are made to the pins of the power cord, but these pins press against certain movable parts which operate the battery switching.

Designs of receivers for battery operation usually necessitate a careful consideration of the electrical and mechanical characteristics of available battery types. Substantial improvement in battery performance and/or life for a given physical size has been made during the years. An American Standard is available[1] giving dimensions, tests, performance, and terminal arrangements.

MINIATURIZATION

130. Methods and Results.[2] Reduced size offers advantages for three-way portable receivers as well as for virtually all other classes of electronic equipment. Means for obtaining reduced size include the use of (1) transistors instead of tubes, (2) miniaturized capacitors, coils, resistors, speakers,[3] and other components, and (3) printed-circuit wiring and other special assembly methods. If these various techniques are intensively applied, a point is likely to be reached where inadequate heat dissipation and consequent excessive temperatures threaten to damage the circuit parts.[4] For dissipations above about 1 watt per sq in. of exposed surface, simple radiation and free convection do not afford sufficient cooling.[5]

A subminiaturized design of an air-borne receiver tunable over the range from 190 to 550 kc, in which tubes were retained but various other special components were incorporated, was made by the National Bureau of Standards.[6] The size was reduced to 55 cu in. (corresponding to a cube 3.8 in. on a side), in which a 12-tube circuit was contained having a sensitivity in the region of 1 to 4 μv, image and i-f rejections exceeding 80 db, and ample output for headphone operation. The i.f. was 135 kc.

A *printed circuit* consists typically of a paper-base phenolic board having on one side copper foil of suitable arrangement to constitute the various connections and on the other side the circuit components, such as resistors, coils, and capacitors, with holes through the board for the leads from the components. With this construction a single dip of the copper-foil side into a solder bath makes all connections to the components.[7] This manner of making the circuit connections has the advantages of compactness, reliability, and adaptability to economical manufacture. The term printed circuit is also applied to ceramic assemblies such as the capacitor and resistors constituting an interstage coupling.[8]

Additional data regarding miniaturization are given in connection with special manufacturing methods in Sec. 203.

Miniaturized i-f amplifiers for radar service, using subminiature tubes, can be designed with a volume, e.g., of 5.25 cu in. for a six-tube unit. With special attention to the various components, ambient temperatures as high as 85 to 100°C, producing

[1] "Specification for Dry Cells and Batteries," Standard C18.1-1954 of American Standards Association and *Nat. Bur. Standards Circ.* 559, available from Government Printing Office, Washington 25, D.C.

[2] Henney, K., *et al.*, eds., "Reliability Factors for Ground Electronic Equipment," McGraw-Hill, 1956; see pp. 6-8 through 6-19 on miniaturization and encapsulation. Jacobs, Y., "Fundamentals of Electronic Equipment Miniaturization," Publication PB-103,547 of U.S. Office of Technical Services, Washington 25, D.C., 1951.

[3] Bleazey, J. C., J. Preston, and E. G. May, Miniature Loudspeakers for Personal Radio Receivers, *RCA Rev.*, **17**, 57, March, 1956.

[4] "Heat Transfer in Miniaturized Electronic Equipment," 1955; available as Publication D 211.2:El 2/2 from Government Printing Office, Washington 25, D.C.

[5] Danko, S. F., and R. A. Gerhold, Printed Circuitry for Transistors, *Proc. IRE*, **40**, 1524, November, 1952.

[6] Shapiro, G., Subminiaturization Techniques for Low-frequency Receivers, *Nat. Bur. Standards Circ.* 545, April, 1954.

[7] "Proceedings of the Symposium on Printed Circuits," Engineering Publishers, 1955. Waverling, E., Printed Circuits for Home Radio Receivers, *Electronics*, **25**, 140, November, 1952. Davis, B. L., Printed Circuit Techniques: An Adhesive Tape-resistor System, *Nat. Bur. Standards Circ.* 530, February, 1952. National Bureau of Standards, "Printed Circuits: Final Report," Publication PB-100,950 of U.S. Office of Technical Services, Washington 25, D.C., 1950. Brunetti, C., and R. W. Curtis, Printed Circuit Techniques, *Proc. IRE*, **36**, 121, January, 1948.

[8] "Unit Standards for Ceramic Based Printed Circuits," EIA Standard RS-161, August, 1956. "Test Standard for Ceramic Based Printed Circuits," EIA Standard RS-162, August, 1956.

internal temperatures up to 175 to 200°C, can be accommodated. Grooving or fins for cooling will ordinarily lower the internal temperature by 10 to 20°C at the cost of additional equipment volume.[1]

REFLEX OPERATION

131. Reflex Principle. Although little used in the United States, the reflex principle is one of the standard methods available to the radio design engineer. A reflex circuit has been defined in the IRE Standards as one through which a signal wave passes both before and after a change of frequency. Since such a circuit has currents of two different frequencies, it has two inputs and two outputs, and filtering is necessary to separate the two outputs.

The field of usefulness is where power levels are low and the number of tubes is restricted by economic or technical factors. European broadcast receivers have often included a reflexed tube serving as both an i-f and a-f amplifier. Both pentodes and converter tubes have been used in this way. In the case of the converter tubes, the mixing feature is not required, but the two input grids offer convenient points for applying the two frequencies. Another possibility is to amplify both the r-f signal as received and the i-f signal, but such designs are likely to have poor performance for radio frequencies which are harmonics of the i.f.

A tube operating in a reflex circuit is subjected to the sum of the peak voltages and currents at both operating frequencies, so that the power levels must not be too high. If care in this regard is not taken, distortion will occur; in particular, tubes handling i-f and a-f signals may give poor audio quality. Besides keeping the power levels low, points worth some consideration are increasing the plate-supply voltage within the rating of the tube, or choosing a larger tube. The power limitation will probably make it undesirable to apply a.v.c. to a reflexed stage; however, if the stage handles i-f and a-f signals and a.v.c. is applied, the control of the a-f gain constitutes forward-acting a.v.c., so that a very flat a-v-c curve should be obtained.

Wide use of the reflex principle has been made in Australia, particularly in receivers having three tubes and a power rectifier, the three tubes serving as (1) converter, (2) reflexed i-f and a-f amplifier and diode detector, and (3) output audio stage.[2]

F-M RECEIVERS

Frequency modulation is extensively used for such purposes as police service around 35 Mc, broadcasting around 100 Mc, automobile and railroad telephone around 155 Mc, and the sound channels of television.[3]

132. Modulation Index. The ratio of maximum frequency swing to one side divided by the highest frequency component in the modulating signal is called the *modulation index*. In police service the deviation is usually 15 kc to each side, and the highest a.f. transmitted is 3 kc, giving a modulation index of 5. For present-day broadcast practice the corresponding figures are 75 and 15 kc, again giving a modulation index of 5. For the sound channel of television broadcasts the maximum deviation is 25 kc.

On certain segments of the amateur bands the FCC has authorized the use of "narrow-band" f.m. which is subject to the requirement that "the peak deviation is limited to a value equal to or less than the maximum modulation frequency." This corresponds to a modulation index of unity or less.[4]

[1] Scal, R. K-F., Miniature Intermediate-frequency Amplifiers, *Nat. Bur. Standards Circ.* 548, July, 1954.

[2] Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 1140–1146 and 1258–1259.

[3] A general discussion and an extensive bibliography on f.m. are given by S. W. Seeley, Frequency Modulation, *RCA Rev.*, April, 1941, p. 468. Hund, August, "Frequency Modulation," McGraw-Hill, 1942. Kiver, Milton S., "F-M Simplified," Van Nostrand, 1947. Tibbs, C. E., with revision by G. G. Johnstone, "Frequency Modulation Engineering," 2d ed., Wiley, 1956. "F-M Transmitters and Receivers," Army Manual TM 11-668, 1952; available from Government Printing Office, Washington 25, D.C.

[4] Miller, E. C., Frequency Discriminator for Narrow-band F-M, *Electronics*, **26**, 128, February, 1953.

133. Transmitter Specifications Affecting F-m Broadcast Receivers. For f-m broadcasting the FCC "Standards of Good Engineering Practice" establishes 88 to 108 Mc as the f-m broadcast band, divided into 100 channels of 200 kc each.

The FCC adopted in 1947 a rule that the normal spacing between stations in a given locality will be 800 kc to avoid interference difficulties found in tests in Syracuse and also to permit the construction of economical receivers having a limited number of tuned circuits for selectivity.

Modulation of 100 per cent in f-m practice does not have a unique definition such as it has in a-m practice. The FCC has standardized this for f-m broadcasting as ± 75 kc. On this basis 30 per cent modulation, which is often used for testing, is a deviation of ± 22.5 kc.

In FCC "Standards of Good Engineering Practice," the signal intensities stated as necessary for satisfactory reception are 1,000 μv per m for city, business, or factory sections and 50 μv per m for rural sections. Objectionable interference is considered to exist for stations on the same frequency if the undesired field strength exceeds $\frac{1}{10}$ of the desired. For stations separated by 200 kc, the corresponding figure is $\frac{1}{2}$. Stations in a given area will normally not be assigned frequencies differing by 10.7 Mc, since this is the usual i.f. for f-m receivers.

The audio characteristic of f-m broadcast transmitters must extend from 50 to 15,000 cycles, with preemphasis on the upper audio frequencies in accordance with Sec. 134. The maximum audio harmonic distortion is specified as 3.5 per cent total harmonics for any fundamental frequency between 50 and 100 cycles, 2.5 per cent for any fundamental from 100 to 7,500 cycles, and 3.0 per cent for any fundamental from 7,500 to 15,000 cycles. The f-m noise level in the transmitter output must be at least 60 db below full modulation; *i.e.*, the noise deviation must not exceed 75,000/1,000, or 75 cycles. The a-m noise level must be at least 50 db down, which for a 1-kw station means less than 0.01 watt.

134. Preemphasis of Upper Audio Frequencies. Since most of the energy in speech and music is contained in the lower audio frequencies, it is possible to reduce the noise in a system by reenforcing the medium and high audio frequencies at the transmitter and then correspondingly attenuating these regions at the receiver. In this way the modulation of the transmitter for these frequencies is increased, requiring a slight reduction of modulation for the lower frequencies, while a large reduction of noise is obtained by means of the compensating attenuation at the receiver. For this reason the FCC specifies that "preemphasis shall be employed in accordance with the impedance-frequency characteristic of a series inductance-resistance circuit having a time constant of 75 μsec." The significance of this statement, which specifies preemphasis in microseconds, can be visualized by assuming that the inductance and resistance are connected in series in the plate circuit of a constant-current audio source, such as a pentode amplifier. At the lowest frequencies the impedance of the inductance will be negligible, so that the plate voltage is merely IR. At the high frequencies the reactance of the coil will greatly exceed the value of the fixed resistor, so that the plate voltage is almost equal to IX_L. The plate voltage, therefore, will gradually increase for a given current as higher audio frequencies are applied. The 75-μsec provision merely means that the ratio L/R must be 75×10^{-6}; *i.e.*, there must be 75 μh per ohm of resistance or the equivalent relationship of 13.3 ohms of resistance per millihenry of inductance.

The extent to which the upper frequencies are reinforced by a circuit of 75 μsec time constant is shown by the solid curve in Fig. 31. Since some tolerance is necessary with actual equipment, the FCC requires each station not to have greater preemphasis than this and allows each to have less preemphasis down to the dashed line in the figure. The following formula can be used for computing preemphasis at any frequency:

$$(\text{db})_p = 20 \log_{10} \sqrt{1 + 0.222 \times 10^{-6} f^2}$$

where f is in cycles. This formula, like the solid curve in Fig. 31, gives the preemphasis at any frequency in comparison with the lowest audio frequencies.

The same amount of preemphasis is used in the sound channels of television broadcast stations.

At the receiver it is necessary to introduce an h-f loss, *i.e.*, a deemphasis, to restore the audio frequencies to their proper balance and to realize the advantage in noise reduction. It is usually the practice in receiver design to observe the amount of loss for the upper frequencies present due to the natural characteristics of various elements and then to add a shunt capacitor at one point of such value as to increase the loss to the necessary amount. As a matter of theoretical interest, it may be noticed that, should all the necessary deemphasis loss be inserted at one point, an *RC* combination

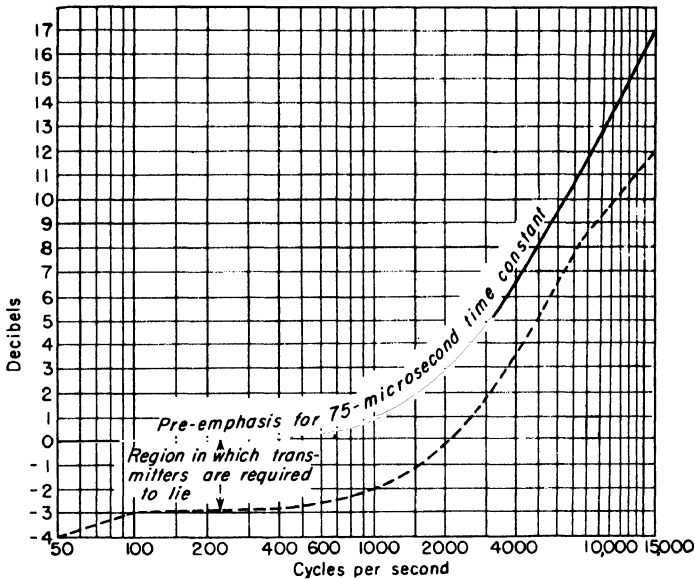



FIG. 31. Preemphasis of upper audio frequencies in f-m broadcasting and in sound channel of television system.

of 75 μsec would be required, such as furnished by 50,000 ohms in parallel with 1,500 μμf or by 25,000 ohms in parallel with 0.003 μf.

A different standard preemphasis is used in f-m mobile services. In this case a uniform rate of 6 db per octave is employed.

135. EIA Standards for F-m Receivers. The EIA has standardized upon a balanced 300-ohm transmission line for the input connection of f-m broadcast receivers. A parallel-wire construction with a continuous low-loss plastic separator is available for this service.

The standard i.f. is 10.7 Mc. This value slightly exceeds half of the 20-Mc total width of the f-m band, which is 88 to 108 Mc, so that no f-m station will lie on the image when another is being received. The possibility, of course, still remains that image interference may be encountered from other services outside the 88- to 108-Mc band.

Under the EIA standards and also the requirements of FCC, f-m broadcast receivers are limited in the allowable amount of oscillator interference, as discussed in Sec. 20. In the case of a receiver with a triode r-f stage and a triode modulator, a reduction of radiation can be obtained by applying the oscillator injection on the cathode and the r-f signal on the control grid of the modulator and using a neutralizing circuit to

annul the oscillator voltage on the control grid of the modulator. Bridge arrangements can also be used for controlling radiation.[1]

Tests to determine the various properties of f-m broadcast receivers are described in Sec. 22. F-m receivers for mobile services are tested as described in Sec. 167.

136. Application of Negative Feedback to Oscillator of F-m Set. If the output audio voltage delivered by an f-m receiver is used to produce f.m. of the local heterodyne oscillator of the receiver with such a direction as to reduce the frequency difference between the oscillator and the incoming signal, a form of negative feedback is obtained[2] in which the frequency swing of the i-f signal is reduced. In this way distor-

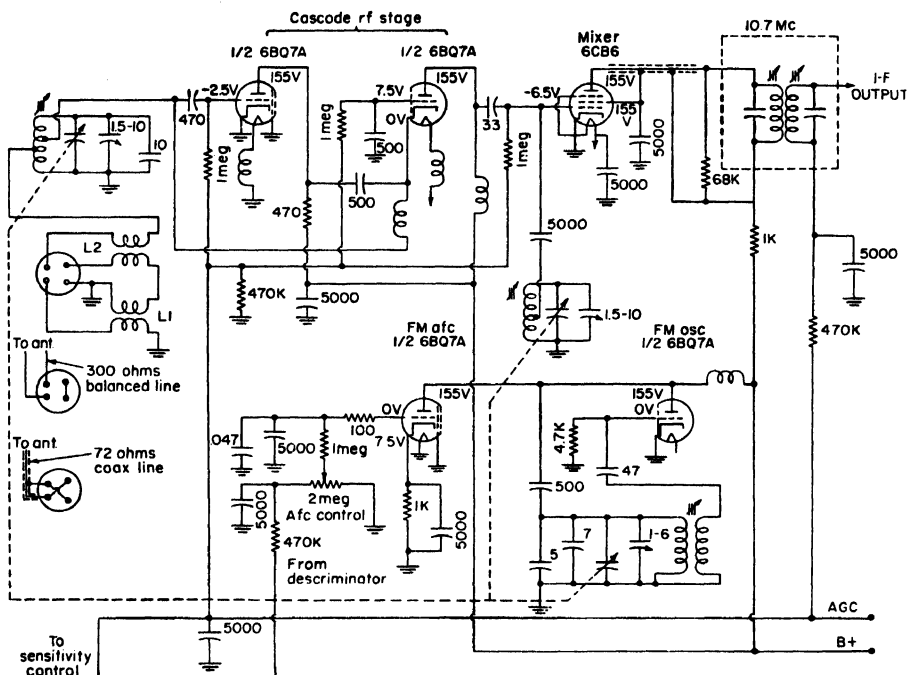



FIG. 32. Front end of high-grade f-m tuner. (*Courtesy of Fisher Radio Corp.*)

tion originating in the receiver can be greatly reduced, so that the method is of special interest in connection with multiplex f-m systems, such as used in point-to-point practice.[3]

137. F-m Broadcast Receivers. Although numerous techniques are available for the design of receivers for f-m broadcasting, experience has shown that acceptable results at the lowest cost are obtained with a rather standardized group of features. These include the superheterodyne circuit tuned by a capacitor gang with the oscillator frequency higher than the signal. An r-f stage is generally used, and the more selective sets have a three-gang capacitor for tuning the antenna circuit, the interstage coupling, and the oscillator. A less selective design is also made in which a two-gang capacitor tunes the interstage coupling and the oscillator; *i.e.*, the antenna circuit has a fixed broad tuning for the entire band from 88 to 108 Mc.

[1] de Quant, H., and P. Zijp, The FM Section of Modern Broadcast Receivers, *Philips Tech. Rev.*, **17** (12), June, 1956.

[2] Chaffee, J. G., The Application of Negative Feedback to Frequency-modulation Systems, *Proc. IRE*, May, 1939, p. 317.

[3] Clavier, A. G., and G. Phelizon, Paris-Montmorency 3000-Mc F-m Link, *Elec. Commun.*, June, 1947, p. 159.

Low-noise types of r-f stages, especially the cascode circuit,[1] are suitable for high-grade f-m tuners, such as offered for high-fidelity receivers. An example of this in a manufactured tuner is shown in Fig. 32. The circuit is seen to include also automatic frequency control (a.f.c.). Another feature is the provision for either a 300-ohm balanced or a 72-ohm coaxial transmission line, bifilar coils L_1 and L_2 serving as balun for the 300-ohm case and as impedance matcher for the 72-ohm case.

A minor disadvantage of capacitive station tuning is that the common coupling in the capacitor rotor shaft in a three-gang design limits the gain of the r-f stage to about 20 db. The capacitive design is, however, generally used for other reasons.

138. Design Points on F-m Broadcast Receivers. The 100-Mc f-m band is high enough for the grid losses of tubes to be important. Tubes suitable for use as r-f amplifiers include the 6AK5, 6AC5, 6SC7, and 6BA6 pentodes and the 6BQ7A and 12AT7 double triodes. For converter, suitable types are the 7F8 and 12AT7 double triodes, the 6CB6 and 6AK5 pentodes, and the 6SB7Y and 6BE6 heptodes. As separate oscillator, if one is used, the 6C4 triode and the 7F8 are suitable.

The i-f and detector design may consist of two i-f stages, followed by a grid-current limiter and a discriminator. Such a limiter, however, does not reject a.m. of fairly high frequency, such as the 400- or 600-kc undulation due to interference from the second or third channel away in frequency allocation; in fact, these limiters increase such interference. This drawback is avoided and one tube saved by using only one i-f stage, followed by a driver stage and a ratio detector.

For the reception of 100-Mc f-m broadcasts in automobiles, an experimental receiver has been used with limiter and f-m detector of substantially greater band widths than in a home receiver. This design was reported to afford good reception despite multipath propagation conditions.[2]

139. Tuning Indicators. High-grade f-m receivers should be provided with a tuning indicator. Medium-price sets can be tuned fairly well by the user without a tuning indicator if (1) the selectivity curve falls off about 6 db at ± 75 kc, (2) no limiters are used, (3) a ratio detector is used, and (4) only a limited amount of a-v-c action is provided.

A simple tuning-indicator circuit for receivers having limiters is one which develops a control voltage from the grid current of a limiter stage.

The 6AL7GT electron-ray tube with three deflecting plates is specially suitable for f-m receivers and is discussed in Sec. 91.

140. F-m Detectors. The phase-variation, discriminator, or "coupled-circuit" type of f-m detector is similar to the discriminator of an a-f-c system described in Sec. 97.[3] While the carrier frequency is not being deviated, this detector rejects a.m. by virtue of the equal and opposite rectified voltages in its output circuit. However, when the frequency is at some deviated value, one rectified voltage exceeds the other, and it can be seen that an increase of level due to a.m. will increase both rectified voltages and therefore their difference. For example, if the voltages were -5 and $+6$, with no a.m., and were doubled by the a.m. to -10 and $+12$, the differential output would be increased from 1 to 2 volts.

The choice of type of f-m detector for a sound broadcast receiver (for 88 to 108 Mc) is more restricted than for the sound channel of a television receiver for the reasons that (1) normally higher audio fidelity is desired, and (2) the variation of the 10.7-Mc carrier when tuning to a station must be accommodated in distinction to the fixed 4.5-Mc carrier in the intercarrier television receiver.

141. Ratio Detectors. Detectors[4] having the property of being insensitive to a.m., while the carrier is deviated as well as when it is not, have been developed. Of special

[1] Wallman, H., A. B. MacNee, and C. P. Gadsden, A Low-noise Amplifier, *Proc. IRE*, **36,** 700, June, 1948.

[2] Onder, K., Mobile F-M Broadcast Receiver Design, *Electronics*, **27,** 130, May, 1954.

[3] Scroggie, M. G., Low-distortion F.M. Discriminator, *Wireless World*, **62,** 158, April, 1956.

[4] Beers, G. L., A Frequency-dividing Locked-in Oscillator Frequency-modulation Receiver, *Proc. IRE*, December, 1944, p. 730. Bradley, W. E., Single-stage F-M Detector, *Electronics*, October, 1946, p. 88. Loughlin, Bernard D., Performance Characteristics of FM Detector Systems, *Tele-Tech*, January, 1948, p. 30.

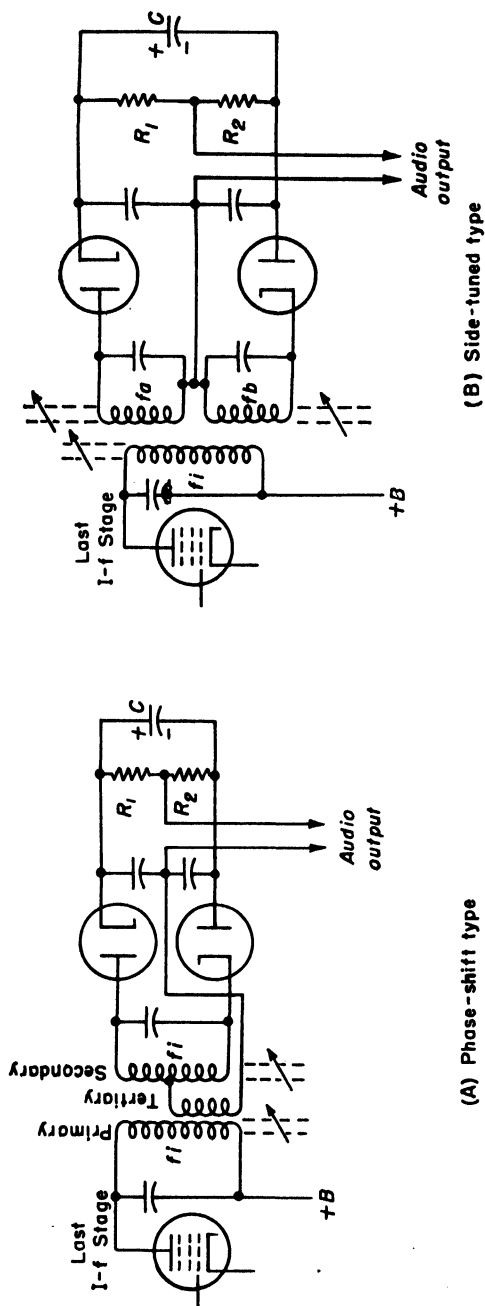



FIG. 33. Essential parts of the two general types of ratio detectors.

interest is the ratio detector,[1] which has this property and obtains its name because of its sensitiveness to the ratio of the two voltages developed from the applied signal, rather than to their difference. Since the ratio between the two voltages is the same at any level, an ideal ratio detector does not respond to a.m. at any frequency deviation.

There are two general types of ratio detectors, the essential parts of which are shown in Fig. 33. The phase-shift type is the more widely used. Fundamentally, the two types differ only in the use of a different i-f coil structure between the last i-f stage and the diodes. In common with discriminators, ratio detectors make use of phase shift or of side tuning to develop the necessary frequency-sensitive i-f voltages. The tuning of the side circuits in the side-tuned type is shown in Fig. 33*B* by the notation f_a and f_b to indicate the resonant frequencies of the two side-tuned circuits, one above and one below the center value of the i.f., f_i.

The ratio detector differs from the discriminator in the use of a large capacitor C which maintains a constant voltage during all a-f variations. The time constant of C, R_1, and R_2 has a typical value of $\frac{1}{10}$ sec. For example, C may be 8 μf and R_1 and R_2 may be 6,800 ohms each.

Considerable care is necessary in the design of the transformers and detail circuits of ratio detectors to secure good results. Accurate balance and other refinements are necessary.

A reasonable sensitivity requirement for a driver i-f stage and ratio detector operating at an i.f. of 10.7 Mc is that a signal level of 0.1 volt rms on the driver

[1] Seeley, S. W., and Jack Avins, The Ratio Detector, *RCA Rev.*, June, 1947, p. 201. This paper is also partly given under title Ratio Detector for FM Signals, *Tele-Tech*, July, 1947, p. 46. Hayes, Albert T., Jr., Improved Type of Ratio Detector, *Tele-Tech*, November, 1947, p. 41.

grid with a 22.5-kc (30 per cent) deviation should produce an audio output of about 0.2 volt rms.

The characteristics of the f-m detector in a receiver have an important effect on the apparent tuning action. For example, if the over-all i-f band width at 6 db down is between 150 and 200 kc, the tuning action will be good if the separation of the detector peaks is 300 to 400 kc (± 150 to ± 200); otherwise a noticeable false response may occur on each side of the correct tuning point. The separation of the detector peaks can be observed by using a battery and potentiometer to hold across C a voltage corresponding to the signal strength of interest and then applying in succession various unmodulated i-f signals over a frequency range extending beyond the peaks in each direction.

Theoretical and experimental work[1] confirms that the rejection of a.m. by a ratio detector is produced in a different way from ordinary limiting, such as accomplished

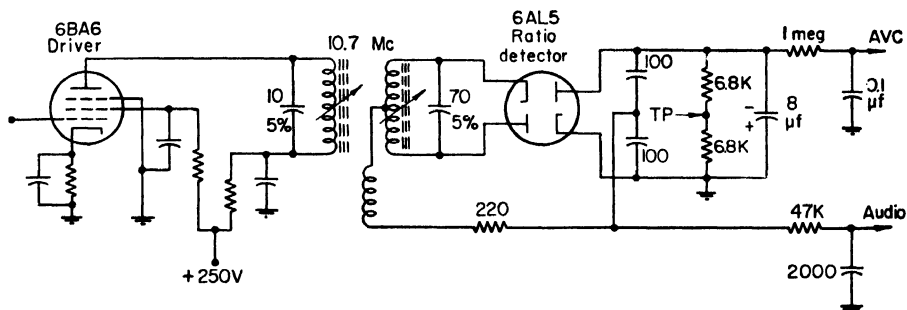



FIG. 34. Practical arrangement of phase-shift ratio detector for 10.7-Mc intermediate frequency. For alignment connect high side of VTVM to audio output and low side to test point marked TP.

by conventional grid-circuit limiters which remove a.m. from the carrier wave. An oscilloscope may show some a.m. on the carrier wave to be removed by a ratio detector; however, this action is incidental and unrelated to the a-m rejection properties. Ratio detectors can be designed to achieve complete a-m rejection when the apparent limiting action is very small. While a diode dynamic limiter (see Sec. 142) has best limiting action with the highest diode conductance, practical ratio-detector circuits usually add resistance in series with each diode to obtain optimum a-m rejection.

Germanium crystals can be used in place of vacuum diodes in ratio detectors, but may require added series resistance to reduce the forward conductance. They also have the limitation that the backward conductance may increase at high signal levels and thereby impair the detector performance.

The grid-to-plate voltage gain of the driver can be substantial. A reasonable maximum figure can be obtained by subtracting 8 or 10 db from the result given by the following expression:

$$20 \log_{10} \left(2 \times \frac{\text{reactance of grid-plate capacitance}}{\text{resistive component of grid impedance}} \right)$$

This formula gives the maximum gain, measured at resonance, for a ratio-detector driver that will just start to oscillate when it is detuned during alignment to the point where maximum feedback occurs. For example, the maximum safe gain at 10.7 Mc for a 6BA6 tube ($C_{gp} = 0.0035\ \mu\mu\text{f}$) in a circuit having an effective grid resistance of 36,800 ohms (Q of 72, C of 30 $\mu\mu$f) would be 46.8 db. The driver gain in this case

[1] Loughlin, B. D., The Theory of Amplitude Modulation Rejection in the Ratio Detector, *Proc. IRE*, **40**, 289, March, 1952.

should not, therefore, be allowed to exceed 37 to 39 db. Note that the value of the resistive component of the resonant impedance in the driver external grid circuit when the primary is well detuned is to be used as the denominator in the formula.

The circuit of a ratio detector with values of the elements is given in Fig. 34.

142. Dynamic Limiters. The grid-bias type of limiter cannot be given a short enough time constant to follow the envelope and eliminate a.m. of high frequency, such as the 400 kc encountered in second channel interference. This difficulty can be avoided by the use of a "dynamic limiter" which utilizes one or two diodes that respond instantly to amplitude variations and limit them. Like a ratio detector, it

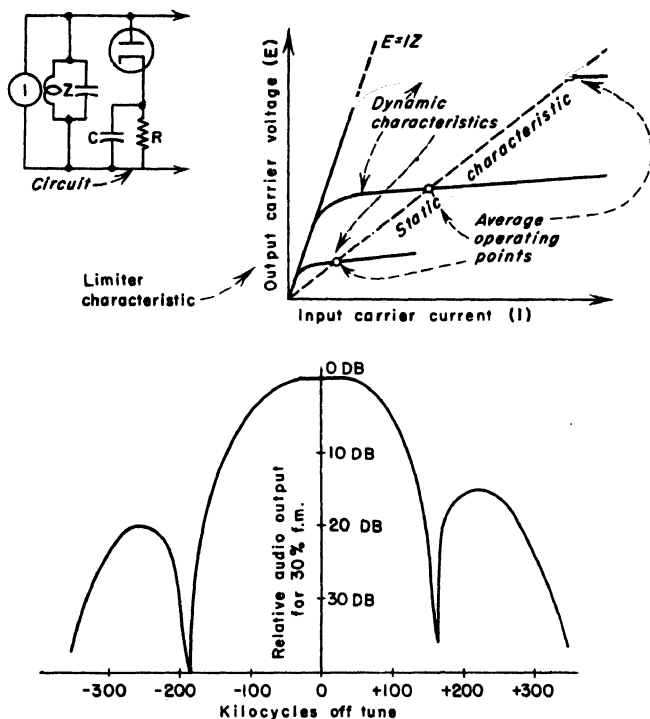



FIG. 35. Dynamic limiter for f-m receiver.

has an *RC* circuit which holds its characteristics constant with regard to a-f variation but permits slow change to adjust the signal level. The dynamic-characteristic curves (Fig. 35) are therefore almost flat, while the static characteristic is a linear output-input relation. Used with a phase-variation discriminator (which is not a ratio detector), the dynamic limiter gives the receiver the tuning action shown in Fig. 35. The side responses are fairly weak, being about 17 db down. There is relatively little interstation noise, a property similar to that of a ratio detector.

Practical further utilization of the dynamic limiter can be made by the substitution of a germanium crystal in place of the vacuum diode. By the use of two diodes or crystals with opposite polarity, both halves of the modulation envelope can be controlled. Also the dynamic limiter can be used with a ratio detector with very good performance and a relaxation of the requirements on each. It is worth noting that the use of a dynamic limiter with a ratio detector affords the best performance now obtainable in an f-m receiver.

HIGH-FIDELITY RECEIVERS

Most radio receivers produce a quality of sound which leaves a good deal to be desired in comparison with the original music or speech. The defects include restriction of the frequency and amplitude ranges and the introduction of hum, noise, and extraneous harmonic and intermodulation frequencies. Receivers which substantially surpass the average set in the quality of reproduction are called "high-fidelity receivers."[1]

143. Characteristics of Ear and of Music, Speech, and Sound Effects. The characteristics of the ear[2] for persons twenty years of age include a frequency range from 20 to 16,000 cps with the greatest sensitivity over the portion from 1,000 to 5,000 cps. In this central range an average twenty-year-old listener can hear sounds over the large intensity range of 105 db at 1,000, 110 db at 2,500, and 105 db at 5,000 cps. For higher or lower frequencies, the intensity range of the ear greatly declines as this central frequency region is departed from. This intensity range is, however, not fully utilizable in listening to the average receiver operating with a loud-speaker because of interfering room noise which is always present and masks faint sounds that would otherwise be heard. The sensitivity of the ear for the higher frequencies declines with increasing age, especially for men so that at age fifty-five the sensitivity at 7,040 cycles is −24 db for men and −19.7 db for women.[3]

Although frequencies as low as 20 cps and components as high as 15,000 cps exist in music, an analysis by Fletcher[4] leads to the conclusion that under practical circumstances substantially complete fidelity for symphonic music is obtained with a volume range of 65 db and a frequency range from 60 to 8,000 cps. For speech the corresponding figures are 40 db and 100 to 7,000 cps.

The single-channel character of the normal radio system produces what may be called a "monaural" type of operation, so that the binaural or "stereoscopic" feature made possible by the two ears of the listener is not used. This is a standard limitation which has been removed in special demonstrations.[5]

144. Preferences of Listeners. There is a difference of opinion as to the reason why a typical listener removes the higher pitches of music with his tone control. A careful analysis of various data and procedures in this connection, published by LeBel,[6] should be studied by all who are concerned about the matter.

A direct acoustic test of listener preference for the upper frequencies was made by Olson,[7] the apparatus including no electrical equipment. Original music and speech with a full frequency range were compared with the same sound except that frequencies above 4,000 cps were removed. An acoustic filter, made up of large panels that could be rotated in and out of position, was used with an opaque sound-transparent screen to hide it from the audience. The room had dimensions 24 ft long by 20 ft wide by 9½ ft high. The musical tests were made with a six-piece orchestra in one corner and the listeners in the opposite corner. For speech tests the orchestra was

[1] Rettinger, M., "Practical Electroacoustics," Chemical Publishing, 1955. Darrell, R. D., Environmental-fitness Considerations of Audio High-fidelity Systems, *IRE Convention Record*, **3**, Part 7, 100, 1955. Newitt, J. H., "High Fidelity Techniques," Rinehart, 1954. Canby, E. T., "Home Music Systems," Harper, 1953. Canby, E. T., C. G. Burke, and I. Kolodin, "The Saturday Review Home Book of Recorded Music and Sound Reproduction," Prentice-Hall, 1952. Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 603–634, 1234–1239, and 1321–1325.

[2] Fletcher, Harvey, Hearing, the Determining Factor for High-fidelity Transmission, *Proc. IRE*, June, 1942, p. 266.

[3] Steinberg, J. C., H. C. Montgomery, and M. B. Gardner, Results of World's Fair Hearing Tests, *Bell System Tech. J.*, October, 1940, p. 533.

[4] *Loc. cit.*

[5] *J. Soc. Motion Picture Engrs.*, October, 1941, p. 331. Tinkham, R. J., Stereophonic Recording Equipment, *Trans. IRE*, **AU-1** (6), 13, November–December, 1953.

[6] LeBel, C. J., Psycho-acoustical Aspects of Listener Preference Tests, *Audio Eng.*, August, 1947, p. 9. See also Morgan, H. K., Natural Sound Reproduction, *Trans. IRE*, **AU-2** (4), 106, July–August, 1954.

[7] Olson, H. F., Frequency Range Preference for Speech and Music, *Electronics*, August, 1947, p. 80; also given under title, Report on Dr. H. F. Olson's Listener Preference Tests, *Audio Eng.*, June, 1947, p. 27.

replaced by one person. Over a thousand listeners participated in the tests and were told that there was a source of music or speech behind the curtain and that the program would be rendered under two different conditions, indicated by the display of letters *A* or *B* on the indicator, and that these conditions would be changed every 15 sec. The listeners were asked to vote on which condition they considered preferable. About two-thirds of the observers preferred the wide range, the figures being about the same for dance music, semiclassical music, and speech. At a meeting of the Acoustical Society of America, J. P. Maxfield, formerly of the Bell Telephone Laboratories, reported that with electronic systems free from h-f transients and cross modulation, he had observed listeners to prefer wide-band reproduction. This, he stated, was true for both single-channel and stereophonic systems.

Extension of the upper audio range greatly increases the requirements for low noise and low introduction of harmonics and intermodulation frequencies. If these faults are present, the upper register will detract from, rather than improve, the over-all quality, and the average listener will unhesitatingly turn down the tone control.

145. Fidelity of A-m and F-m Receivers. The 10-kc spacing of stations in the standard broadcast band places a definite upper limit on the reproduction of the higher audio frequencies. This is 4,000 or 5,000 cps for distant reception and may be substantially higher for strong local signals.

When one end of the audio range of the receiver is restricted, the most natural results are obtained by proportional restriction at the other end of the frequency range; *i.e.*, the l-f range should be extended downward in the same relative proportion as the upward extension of the h-f range. A common working rule is that the product of the lowest and highest frequency limits should be approximately 500,000. Table 13 gives typical ranges[1] and the corresponding frequency product for four types of receivers.

Table 13. Audio-frequency Ranges of Typical Receivers*

Type	Lower frequency limit	Upper frequency limit	Frequency product
Small table set	190	2,700	514,000
Large table set	140	3,500	490,000
Typical console	100	4,500	450,000
High-fidelity console	80	6,600	528,000

* Hanson, O. B., Comments on High Fidelity, *Electronics*, **17**, 130, August, 1944.

If the frequency range goes above 8,000 cycles, an a-m receiver is likely to bring in "monkey chatter," produced by undesired side frequencies beating with the desired carrier or by an undesired carrier beating with the desired side frequencies. It is only with very strong local signals that high-fidelity reception above 8,000 cycles can be enjoyed. In this case the next hurdle is the 10-kc whistle between carriers on adjacent channels; this can be removed with a band-elimination filter.

To obtain a receiver that will produce fairly high-fidelity output on local stations and still be useful for the reception of distant signals, variable selectivity should be provided.

An a-m receiver intended for high-fidelity reception should have special attention devoted to the last i-f amplifier to avoid overloading at this point. It may be necessary to reduce or remove the a-v-c bias on this stage. The audio signal is, of course, the envelope of the r-f wave, so that overloading, which would distort the envelope, must be avoided.

Another refinement is approximately equal audio and d-c loads for the diode detector. Failure to provide this equality will cause clipping of the output during the

[1] Hanson, O. B., Comments on High Fidelity, *Electronics*, **17**, 130, August, 1944. Clark, W. B., Over-all Audio Performance of Commercial Radio Receivers, *Trans. IRE*, **PGA-7**, 11, May, 1952.

inward modulation (also called the "downward" modulation), so that distortion in the reproduced audio wave will occur, especially for modulation exceeding about 50 per cent.

The f-m transmission standards permit the reception of high-fidelity f-m programs with reasonable care. The design must include provisions for a-m rejection to remove noise and to remove distortion which otherwise would result from a.m. introduced by the rounding of the tops of the selectivity curves of the i-f transformers. Beyond this, the requirements for high fidelity involve only the problems of audio amplification and the loud-speaker.

146. General Requirements for High-fidelity Set. The chief general requirements to be satisfied are as follows:

1. The set must be practically free of hum.
2. The reproduction must be practically free of harmonics and intermodulation frequencies at all volume levels.
3. The electrical characteristics must be practically uniform over a wide audio range.
4. A good loud-speaker must be provided.

Interesting general treatments of these requirements[1] and of corresponding studio and motion-picture techniques[2] may be consulted in the literature.

A critical portion of an audio amplifier from the fidelity standpoint is the output stage because of the high power level.[3] This stage is normally operated push pull to remove the even harmonics and to remove the d-c flux in the output transformer.[4] There is a difference of opinion as to the best general type of tube for the output stage. Some favor triodes, owing to the absence of high-order distortion. Others prefer beam tetrodes, with the addition of negative feedback. If triodes are used, care is needed to avoid overloading of the driver stage that furnishes the input for the final stage.

Whatever tubes are used in the last stage, negative feedback is available as a means of reducing distortion. By this powerful tool quality can be improved with little circuit complication, the only important drawback being the reduction in the gain of the amplifier, which can be restored by higher gain couplings or an additional stage.[5]

147. Intermodulation Distortion. Output-input characteristic can be expressed as a mathematical power series in terms of the first, second, third power, etc., giving the instantaneous output as a function of the instantaneous input signal. The second-power term produces the second harmonic and the sum and difference of the various applied frequencies; the third-power term produces the third harmonic and with two input frequencies also produces outputs of $2f_1 \pm f_2$ and $2f_2 \pm f_1$. Counting the two harmonics, the result of third-power distortion is, therefore, six undesired frequencies. Similarly, in the case of fifth-power distortion, two input frequencies produce their

[1] Toth, Emerick, High-fidelity Reproduction of Music, *Electronics*, June, 1947, p. 108. Minter, Jerry, Audio Distortion in Radio Reception, *Proc. Radio Club of America*, January, 1946, pp. 1–5; also available as Chap. 6 of "Standard F-M Handbook," FM Company, Great Barrington, Mass. Hanson, *loc. cit.*

[2] Gurin, H. M., and G. M. Nixon, A Review of Criteria for Broadcast Studio Design, *J. Acoust. Soc. Am.*, May, 1947, p. 404. Maxfield, J. P., Liveness in Broadcasting, *West. Elec. Oscillator*, January, 1947, p. 3. (In Fig. 4 of this paper the largest three ordinates should be 1.0, 2, and 3 rather than 1.0, 1.2, and 1.3). Content, Edward J., The Acoustical Design of FM Studios, *Tele-Tech*, April, 1947, p. 30.

[3] "IRE Standards on Audio Systems and Components: Method of Measurement of Gain, Amplification, Loss, Attenuation, and Amplitude-Frequency Response, 1956," Standard 56 IRE 3 S1, *Proc. IRE*, **44**, 668, May, 1956. Corderman, S. A., and F. H. McIntosh, A New 30 Watt Power Amplifier, *J. Audio Eng. Soc.*, **1** (4), October, 1953. Drisko, B. B., and R. D. Darrell, 40-Db Feedback Audio Amplifier, *Electronics*, **25**, 130, March, 1952. Dammers, B. G., *et al.*, "Application of the Electronic Valve in Radio Receivers and Amplifiers," Book 5 of Philips Technical Series, Elsevier, 1951. Williamson, D. T. N., "The Williamson Amplifier," Iliffe and Sons, London, 1950. Keroes, H., and D. Hafler, Ultra Linear Operation of Williamson Amplifier, *Audio Eng.*, **36**, 26, June, 1952.

[4] Kraus, H. L., Class-A Push-pull Amplifier Theory, *Proc. IRE*, January, 1948, p. 50. On output stages without output transformers see Chap. 15 of this handbook.

[5] Audio Design Notes: Negative Feedback Circuits, *Audio Eng.*, May, 1947, pp. 35, 52. Bode, H. W., "Network Analysis and Feedback Amplifier Design," Van Nostrand, 1945. This book gives a thorough analysis of feedback amplifiers and networks.

fifth harmonics and eight intermodulation products. These frequencies resulting from intermodulation are generally inharmonic or discordant and also likely to fall in the middle or upper portion of the audio range, both of which conditions make them quite objectionable. Intermodulation can be quickly measured with suitable equipment.[1]

148. Audio Transformers for High-fidelity Sets. The output transformer coupling the last tube to the speaker is one of the worst offenders from the standpoint of fidelity. For adequate linearity to avoid intermodulation, much more iron is needed in the magnetic circuit. Both the quality and the quantity of iron used must, of course, be considered.

149. Dividing Networks. Conversion of the a-f signals into sound is facilitated by dividing the frequency range into a lower and an upper part with a separate special speaker for each portion. In this case the h-f speaker is often called a "tweeter," and the l-f speaker a "woofer." A filter network[2] is generally provided to route the proper frequencies to each speaker.

Numerous circuit arrangements are available for dividing networks, such as the mere use of series capacitors in the tweeter circuit. More elaborate networks include two filters which may be designed for series or parallel input connection. For the more elaborate designs a choice of the dividing frequency is required, *i.e.*, the frequency at which an equal electrical output is delivered to the two speakers. Above this frequency, the woofer receives decreasing amounts of signal, and below this frequency the tweeter receives decreasing amounts. The networks used in motion-picture practice generally introduce an attenuation of 12 db per octave as the frequency departs from the dividing frequency. Some experimenters favor locating the dividing network between the driver and the last stage, so as to operate at a lower power level. In this case separate output stages are necessary for the two frequency ranges.

150. High-fidelity Loud-speakers. Acoustics and loud-speakers are treated in Chap. 11; only a few points are discussed here.[3] The tweeter is often mounted in the opening of the woofer, a coaxial arrangement with the advantages of compactness and identity of source location for the two ranges of frequencies.

Diffusing vanes for a cone-type tweeter are desirable to correct the beamlike character of the h-f radiation. An example of a coaxial speaker with such vanes is the Altec-Lansing Duplex.[4] This speaker is used with a dividing network having a crossover frequency of 1,200 cps. With a baffle of 6 cu ft volume, good response is obtained down to 60 cps and with 9 cu ft down to 40 cps. The inner surface of the baffle must be covered with sound-absorbent material to prevent reflections that would give a "hang-over" or echo effect. The vanes give the h-f radiation a distribution of 60 deg horizontally by 40 deg vertically.

Electrostatic speaker units of the push-pull type[5] have attractive characteristics for consideration with regard to the upper and medium portions of the audio range. The push-pull feature, which is also called *balanced operation*, offers the important advantage of freedom from second-harmonic distortion. Other advantages of electrostatic types are smooth response-vs.-frequency characteristic, good spatial distribution, and low transient distortion.

[1] Hilliard, John K., Intermodulation Testing, *Electronics*, July, 1946, p. 123. Pickering, Norman C., Measuring Audio Intermodulation, *Electronic Inds.*, June, 1946, p. 56. Daniel, George, Instrument for Intermodulation Measurements, *Electronics*, March, 1948, p. 134. LeBel, C. J., An Experimental Study of Distortion, *J. Audio Eng. Soc.*, **2**, 215, October, 1954. Wind, M., ed., "Handbook of Electronic Measurements," Polytechnic Institute of Brooklyn, 1956; see Chap. 13 by S. Greenblatt on measurement of distortion.

[2] Schuler, E. R., Design of Loudspeaker Dividing Networks, *Electronics*, February, 1948, p. 124. McProud, C. G., Design and Construction of Practical Dividing Networks, *Audio Eng.*, June, 1947, p. 15. Klipsch, Paul W., Woofer-Tweeter Crossover Network, *Electronics*, November, 1945, p. 144. Briggs, G. A., "Sound Reproduction," 3d ed., Wharfdale Wireless Works, Bradford, Yorkshire, England, 1953; see pp. 190-211.

[3] Briggs, G. A., "Sound Reproduction," 3d ed., Wharfdale Wireless Works, Bradford, Yorkshire, England, 1953; see pp. 14-225. McIntosh, F. H., Speakers and Transmission of Sound Waves, *Trans. IRE*, **PGA-9**, 6, September-October, 1952.

[4] Lansing, James B., The Duplex Loudspeaker, *J. Soc. Motion Picture Engrs.*, September, 1944, p. 168.

[5] Leak, H. J., and A. B. Sarkar, Full-range Electrostatic Loudspeakers, *Wireless World*, **62**, 486, October, 1956.

VARIABLE SELECTIVITY

151. Types and Properties. Ability to vary the selectivity of a receiver is useful where wide variations in receiving conditions are encountered. In broadcast reception, variable selectivity permits the use of a wide carrier-frequency pass band for the high-fidelity reception of local programs, while upon contraction of the selectivity distant stations can be received free of cross talk. In amateur practice, very sharp selectivity is often required to receive phone and telegraph messages in the presence of intense interference. Amateur receivers generally include variable selectivity[1] with an adjustable quartz-crystal filter. By this means band widths as small as 100 cycles can be obtained.

Variable selectivity is easily obtained in the superheterodyne receiver by a change of the characteristics of the i-f amplifier. Available methods are (1) variation of the coupling between the windings of one or more of the i-f transformers so as to widen the selectivity through over-optimum coupling;[2] (2) introduction of damping, *i.e.*, lowering the Q of one or more of the coupling systems (this has the disadvantage of reducing the selectivity against interference having a frequency some distance from the carrier frequency); (3) detuning the i-f tuned circuits, so as to convert the amplifier to stagger-tuning; and (4) use of an auxiliary feedback tube which can be made to cut down the peak and build up the skirts of the selectivity curve.

Some methods produce symmetrical variation of the selectivity while others produce unsymmetrical variation, the distinction being whether the selectivity curve for the various adjustments is the same or different on the two sides of resonance. In Table 14 the chief characteristics of the two types of operation are given.

Table 14. Characteristics of Symmetrical and Unsymmetrical Types of Curves for Variable Selectivity*

Characteristic	Symmetrical type	Unsymmetrical type
Complication of operation	Only one variable, the amount of expansion	Two variables, *i.e.*, the response on the two sides
Fidelity	Not critical	May be critical because of carrier being near edge of band
Selectivity	Better when comparable interference is encountered on both sides of the carrier	Better when major interference is on only one side
Noise	Minimum	About 3 db above minimum
General usefulness	Superior from fidelity standpoint	Superior from selectivity standpoint

* From Wheeler, H. A. and J. K. Johnson, High Fidelity Receivers with Expanding Selectors, *Proc. IRE*, June, 1935, p. 594.

A laboratory receiver with a continuous manual variation of selectivity was constructed at Hazeltine and is described in a paper[3] by Wheeler and Johnson. This design uses variation of coupling in two i-f transformers, of equal Q, and one additional interstage coupling of $Q/2$ to obtain an approximately flat-topped selectivity curve. The valley between peaks caused by the overcritical coupling of the adjustable transformers is filled in by the single peak of the $Q/2$ transformer. Broadcast receivers with variable selectivity generally provide two or three band widths. One or two small windings are used in each variable transformer to increase the coupling to the desired overcritical value. When only the regular windings are in use, the transformer has critical or slightly undercritical coupling, and the selectivity is sharpest.

[1] Villard, O. G., Jr., and W. L. Rorden, Flexible Selectivity for Communications Receivers, *Electronics*, **25**, 138, April, 1952.

[2] Sturley, K. R., "Radio Receiver Design: Part I," 2d ed., Wiley, 1953; see pp. 517-550.

[3] Wheeler, H. A., and J. K. Johnson, High Fidelity Receivers with Expanding Selectors, *Proc. IRE*, June, 1935, p. 594.

Laboratory receivers have been made in which the selectivity curve automatically expands as stronger signals are tuned in, and contracts as weaker signals are received. In a design described by G. L. Beers[1] triode tubes reduce the Q of various windings of the coupling transformers.

A discussion of automatic selectivity control by H. F. Mayer[2] compares variation of coupling, variation of Q, and detuning, and leads to the conclusion that alteration of the coupling produces the best results for the additional parts required.

By means of an audio tuned circuit, it is possible to isolate the 10-kc beat between the carriers of adjacent broadcast stations and utilize this to increase the selectivity, thus making the receiver selective to a degree depending upon the strength of the interference.[3] This design includes a control tube for each of two i-f stages, the control tube feeding back energy to reduce the peak of the selectivity curve by degeneration and build up the sides by regeneration. For weak signals or strong 10-kc beat, the control tubes are biased toward cutoff. As stronger signals are received, or weaker interference, this negative bias is reduced, thus widening the selectivity.

Another automatic selectivity control[4] includes two traps tuned respectively to the next channel above and the next channel below the carrier frequency. They operate independently to contract the selectivity on the side where interference is experienced. In this way the contraction may be on either side or on both sides.

SINGLE-SIDE-BAND RECEIVERS

152. Advantages and Techniques.[5] The spectrum conservation and other advantages of the single-side-band method led the FCC in 1956 to begin proceedings looking toward the establishment of this feature as a requirement for radiotelephone operation in bands below 25 Mc.

Reception of an SSB transmission is possible in a straightforward manner by the provision of a stable oscillator to simulate the missing carrier. Another method, widely used, depends on phasing operations. In this method the incoming single side band, normally at i.f., is heterodyned separately in two demodulators, one supplied with local injection at reference phase and the other with injection at 90 deg; then the two audio outputs are shifted one ahead 45 deg and the other back 45 deg, with the result that addition of the two audio signals will give the desired signal if the transmission is an upper side band and the difference will give the desired signal if the transmission is a lower side band.

SSB technique has long been used in transoceanic radiotelephony, and advanced methods are utilized to obtain continuity of service and maximum use of the available spectrum.[6] For other services the problem of *compatibility*, that is, operation with both conventional and SSB equipment, may require attention, and it may be desired that a receiver accommodate more than one type of signal.[7]

Transmission with one side band and an amount of carrier inversely proportional to the amplitude of the side band has been successfully demonstrated with experimental equipment in service between airplanes and ground. With this arrangement the full power of the transmitter is devoted to the side band when needed, but at other times

[1] Beers, G. L., Automatic Selectivity Control, *Proc. IRE*, December, 1935, p. 1425.

[2] Mayer, H. F., Automatic Selectivity Control, *Electronics*, December, 1936, p. 32.

[3] Farrington, John F., Receiver with Automatic Selectivity Control Responsive to Interference, *Proc. IRE*, April, 1939, p. 239.

[4] Rust, N. M., O. E. Keall, J. F. Ramsay, and K. R. Sturley, Broadcast Receivers; A Review, *J. IEE (London)*, June, 1941, p. 79; see Sec. 2.6 of the paper.

[5] Approximately 33 technical papers on single-side-band operation are given in *Proc. IRE*, **44**, 1665, December, 1956. Brunette, G. E., and D. E. LeBrun, An Air Force Single Sideband System, *Trans. IRE*, **CS-4** (2), 85, May, 1956. Pappenfus, E. W., An Airborne Single Sideband Transceiver, *Trans. IRE*, **CS-4** (2), 94, May, 1956.

[6] Lowry, W. R. H., and W. N. Genna, An Independent-sideband Short-wave Radio Receiver, *P.O. Elec. Eng. J. (England)*, **46**, 19, April, 1953.

[7] Varian, S. B., Converter for Twin Single-sideband, *Electronics*, **29**, 214, March, 1956.

is more or less used for sending carrier which is available for operating a "mechanical" a-f-c system.[1]

For mobile land service at 160 Mc with channels of the order of 15 to 30 kc, phase-modulation technique appears to afford more service from a limited number of channels than would SSB suppressed-carrier technique.[2]

AUTOMOBILE RECEIVERS

153. Characteristics. The automobile set differs from home sets in that:

1. A small whip antenna is used.
2. Limited space in the car requires compact construction.
3. Vibration and wide temperature range necessitate rugged parts and construction.
4. Inductive tuning of the r-f and oscillator circuits is generally used.
5. The 6- or 12-volt storage battery serves as power supply.
6. Severe disturbance of the engine ignition requires thorough shielding and by-passing.
7. Metal housing is used, generally single-unit.

The conventional automobile-receiver design includes one r-f stage, inductive three-gang tuner for the antenna, interstage, and oscillator circuits, a.v.c., one i-f

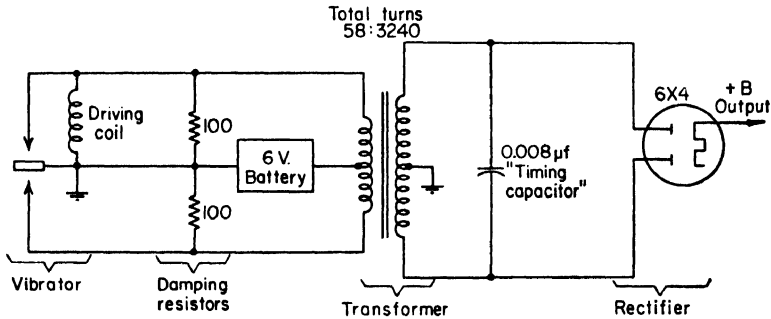



FIG. 36. Typical shunt-coil type of vibrator with stepup transformer and rectifier tube for furnishing plate-supply voltage. The battery can be connected with either polarity.

stage operating at either 455 or 262 kc, a tone control, permanent-magnet dynamic speaker, and power supply consisting of vibrator, transformer, and tube rectifier.[3] Push buttons for choosing preset stations are often provided, the most frequent number being five. These usually operate mechanically to set the inductive tuning gang at the desired position.

Antennas are generally short whip types of low capacitance and limited signal-collecting ability. Dummy antennas to represent these vary from single series capacitors of 40 to 100 $\mu\mu$f to series-shunt capacitor arrangements. In the latter case the series capacitor is connected to the signal generator and the shunt capacitor is connected across the input terminals of the receiver. The values of the capacitor elements lie in the range from 30 to 60 $\mu\mu$f.

The vibrator[4] in the plate-supply circuit commutates the d.c. from the battery so that flux in the transformer reverses direction in step with the moving reed. The frequency is standardized at 115 ± 7 cps. The transformer steps up the voltage, which is then rectified to serve as the desired plate supply. Typical connections for operation with a 6-volt battery are shown in Fig. 36.

[1] Barnes, G. W., A Single-sideband Controlled-carrier System for Aircraft Communication, *Proc. IEE*, **101**, Part III, 121, May, 1954; 399, November, 1954.

[2] Macdonald, A. A., Comparison of Split Channel FM and Single Sideband for Land Mobile Services, *Trans. IRE*, **PGVC-7**, 46, December, 1956.

[3] Langford-Smith, F., "Radiotron Designer's Handbook," RCA, Harrison, N.J., 1952; see pp. 1202–1212, 1275–1278, and 1481.

[4] "Vibrators for Auto Radio," EIA Standard REC-113-C, 1954.

Search tuning for an automobile receiver can be provided without a.f.c. by means of the following mechanical and electronic features: (1) a spring-operated or motor-operated mechanism which can move the normal station-selecting carriage and which if not interrupted will scan the entire broadcast band in about 5 sec; (2) a trigger circuit associated with the a.v.c. in such a way as to stop the sweeping action when a signal of adequate strength is received; (3) a push button or bar which the user can press, thus instructing the receiver to discontinue receiving the particular station and search for another station; and (4) an automatic end-of-the-band return switch whereby the carriage is rapidly sent to the other end of the band whenever needed.[1] This type of operation is quite useful for a driver visiting different areas because the receiver will quickly bring in the various stations which are available in each particular area. The feature can, of course, be provided in addition to the usual push buttons which are normally set to stations in the home territory.

The use of one or two transistors as the output stage of an automobile receiver is discussed in Sec. 38. Designs with this feature can utilize the line of tubes requiring only 12 volts plate supply and thus, with a 12-volt automobile battery, eliminate the vibrator, power transformer, and rectifier of the previously conventional receivers. This line of tubes[2] includes the 12K5, a space-charge tetrode which can deliver the appreciable driver power required by the output stage. Another tube for this service is the 12J8.

The lower-power stages of such an automobile receiver, operating with 12 volts as both heater and plate supply, can utilize the 12AC6, 12AF6, or 12CX6 pentodes in the r-f stage, the 12AD6 heptode converter, various types including the 12CN5 for i-f stages, and the 12AE6, 12F8, or the 12CR6 for detection, a.v.c., and first audio amplifier. A caution has been issued that these tubes should not be tested on the usual emission type of tube checker because of the danger of permanent change of contact-potential conditions.

For a discussion of phonograph operation in automobiles see Sec. 117.

V-H-F AND HIGHER BAND RECEIVERS[3]

154. General Considerations. As receiver designs are made for the v-h-f range and above, various differences in the general conditions which prevail are found. Specially noticeable are (1) the increase of tube input conductance, (2) the smaller amount of atmospheric disturbance and of man-made interference, and (3) the desirability of handling greater band widths. The increase of the input conductance causes lower voltage gains in the coupling circuits. The reduced atmospherics and interference make weaker signals potentially usable. The greater band widths mean that sources of noise in the resistors and tubes of the receiver will produce greater noise power. The result of all these factors is that noise originating in the receiver is a major problem; in fact the design of front ends for receivers for these frequency ranges is largely a matter of obtaining adequate signal-to-noise ratio.[4] Representative noise factors for receivers specially designed for good performance in this respect are 4 db at frequencies up to 100 Mc, 9 db at 1,000 Mc, and 14.0 db at 10,000 Mc.

[1] Hsu, C. C., On the Principle and Design of a Trigger Circuit of a Signal-seeking Radio Using "Difference-Voltage," *Proc. IRE*, **43**, 1591, November, 1955. Guyton, J. H., A Signal-seeking Automobile Receiver, *Electronics*, **26**, 154, May, 1953.

[2] Atkins, C. E., Low Plate-potential Tubes, *Radio and Television News*, **57**, 47, January, 1957. Tube Drives Transistor Output Stage, *Electronics*, **29**, 178, June, 1956.

[3] As general references, see the following: Staff of Radio Research Laboratory, Harvard University, "Very High Frequency Techniques," McGraw-Hill, 1947; T. M. Davis and Emerick Toth, Design of Communication Receivers for the Naval Service with Particular Consideration to the Very-high-frequency and Ultra-high-frequency Ranges, *Proc. IRE*, February, 1947, p. 201; M. J. O. Strutt, "Modern Ultra- and Extreme-shortwave Reception," Van Nostrand, 1947; S. N. Van Voorhis, "Microwave Receivers," McGraw-Hill, 1948; M. S. Kiver, "Introduction to UHF Circuits and Components," Van Nostrand, 1955, with treatment of receivers on pp. 332–384; "The Radio Amateur's Handbook," 35th ed., American Radio Relay League, West Hartford, Conn., 1958, with receiver material on pp. 397–417 and 79–81.

[4] Pritchard, W. L., and K. T. Larkin, The Influence of Noisy Components on the Sensitivity of Microwave Receivers, *Trans. IRE*, **ED-1** (4), 22, December, 1954.

At increasingly higher frequencies, the input conductance of tubes, when operated in the grounded-cathode manner, becomes greater as a result of two causes. One is the presence of inductance in the cathode lead which is common to the input and output circuits of the tube. This is often minimized by the provision of two separate leads from the cathode to separate pins in the base of the tube, one of these being used for the input circuit and the other for the output circuit. Another cause of increasing input conductance is the transit-time effect resulting from the fact that the signal potential on the control grid changes appreciably in the time required for electrons to go from the cathode space charge to the grid, resulting in the signal on the grid doing work on the electrons and thus abstracting energy from the source of the signal and increasing the input conductance of the tube. The conductance due to both the lead-inductance and transit-time effects is proportional to g_m and to f^2. Tripling the frequency causes the input conductance to increase to nine times its previous value.

Since the noise power produced by tubes increases with the band width, the narrower the band, the higher the frequency limit at which any particular type of tube will work. A given tube will work satisfactorily as an oscillator at frequencies considerably above its limit as an r-f amplifier. For amplifier service there is a constant product of the obtainable values of gain and band width, as discussed further in Sec. 194.

Superregenerative detectors for operation at 30 to 3,000 Mc have proved serviceable in various military equipment. For further data see Sec. 159 *et seq.*

Receivers for the frequency ranges now being considered have i.f.s from 5 to 200 Mc.[1] Values of 30 and 60 Mc are widely used in radars. The i-f band widths in various services have values up to 25 Mc. The wider band widths are obtained by staggered tuning or by the use of special feedback arrangements (see Sec. 77).

Superheterodyne frequency converters for receivers operating in these frequency ranges are discussed in Secs. 55 to 57.

155. Use of Lumped-constant Circuits. It is often found that satisfactory tuning control can be obtained by switching in fixed wide-band antenna circuits and r-f tuned circuits upon change of band and varying only the oscillator frequency in tuning from station to station.[2] In a receiver having a band of relatively wide frequency coverage and having low i.f., this technique would lead to objectionable image interference. However, if the limitations are borne in mind, this technique may be found a valuable method. It has been widely employed.

The circuit between an antenna and a conventional grounded-cathode r-f stage can advantageously be given a slight overcoupling in the interest of improved signal-to-noise ratio. This increases the skirt selectivity and thereby removes some noise in the outer frequency regions where no signal components exist.

When conventional coils and capacitors are applied at higher frequencies, a point is reached where the capacitance has been reduced to only that inherently present in the tube, socket, wiring, and coil, and the inductance is as small as possible for the required L/C ratio. Upon replacement of the tube, therefore, there will generally be a different input capacitance, the difference being a large enough fraction of the total circuit capacitance to render inaccurate a calibrated tuning dial or to disturb a unicontrol arrangement. Reduction of the tuning capacitance to where it consists only of the inherent minimums also has the objection that if a.v.c. is applied to the stage, the resulting variation of tube input capacitance may cause excessive detuning.

Another available design for frequencies up to 200 or 300 Mc is the grounded-cathode triode in a push-pull circuit. Laboratory work by William F. Bailey (Hazeltine) has shown that a 6J6 double triode can be cross-neutralized in the conventional transmitter fashion and good gain and signal-to-noise ratio thus obtained. However, the method has the drawback that the stage may require readjustment of the neutralization when a tube replacement is made.

[1] Lebenbaum, Matthew T., Wideband I-f Amplifiers above 100 Mc, *Electronics*, April, 1947, p. 138.

[2] As an example in amateur practice, see B. C. Barbee, A Compact and Inexpensive Superhet for 144 Mc, *QST*, October, 1947, p. 33.

156. Triodes in Grounded-grid Circuit. For signals from about 100 Mc, advantages may be found in the grounded-grid triode[1] as an r-f amplifier, especially if the receiver is intended for wide-band operation. In the grounded-grid amplifier the high side of the input connection is made to the cathode and the low side to the grid, which is grounded. The output is taken between plate and ground in the normal manner. Analysis shows that the input and the output currents are about equal, so that the amplifying action of the tube can be considered as reproducing the input current but at a higher voltage and out of a higher impedance.

The gain of a grounded-grid tube is

$$\text{Voltage gain} = \frac{R_L(\mu + 1)}{R_p + R_L}$$

where R_L = load resistance, ohms
μ = amplification factor of tube
R_p = internal plate resistance of tube, ohms

If the load resistance is much less than the internal tube resistance and if μ is much greater than unity, the gain simplifies to $R_L g_m$.

The value of the grounded-grid stage lies essentially in the fact that the input and the output circuits are fairly well shielded from each other, as in a screen-grid tube, while at the same time the low noise level characteristic of a triode is retained. Another advantage, which is important, is that the grounded-grid stage has an input conductance which is fairly constant with frequency, being

$$g_i = \frac{\mu + 1}{R_p + R_L}$$

If μ is much greater than unity and if the load resistance R_L matches the internal tube resistance R_p, the input conductance is $\mu/2R_p$, or $g_m/2$. If the load resistance is small compared with R_p and μ remains much greater than unity, the input conductance is μ/R_p, or g_m. Since the input conductance of a pentode increases steadily with frequency, there is a frequency above which the grounded-grid triode has lower input conductance than the pentode.

Slightly less than critical coupling into a grounded-grid stage gives the best signal-to-noise ratio.

The 6J4 miniature triode is useful for grounded-grid operation and is rated up to 500 Mc for this service. It has a low value of plate-to-cathode capacitance so as to reduce the feedback from output to input circuit inside the tube. A striking application of this tube was made in the radar receiver used to obtain echoes from the moon at 111.5 Mc. This receiver achieved a noise figure of only 5 db for the front end of the set. For further data on this receiver see Sec. 46.

For tuning in the general region of 300 to 1,300 Mc, especially over limited ranges, transmission lines[2] can be used. A $\lambda/4$ line shorted at the far end acts as a parallel tuned circuit at the near end. Other lengths of shorted line are available for use as high-Q inductors or capacitors. Carefully made transmission lines can be constructed to have the desired reactive characteristics and to have low losses. Balanced parallel-rod lines can be used up to 600 or 700 Mc, and coaxial types above this.

As signals above 400 or 500 Mc are to be received, it becomes advantageous to change from the conventional type of tube to the planar type, and usually at the same time to employ a coaxial type of transmission line for tuning rather than the parallel-rod type. The planar tubes are so-called because the active surfaces of the elements are planes, and include the "lighthouse" series (such as the 2C40)[3] and the Western Electric Type 416-B.[4] For amplification up to 1,000 Mc and oscillation

[1] Jones, M. C., Grounded-grid Radio-frequency Voltage Amplifiers, *Proc. IRE*, July 1944, p. 423. Dishal, M., Theoretical Gain and Signal-to-noise Ratio of the Grounded-grid Amplifier at Ultra-high Frequencies, *Proc. IRE*, May, 1944, p. 276. Murakami, T., A Study of Grounded-grid, Ultra-high-frequency Amplifiers, *RCA Rev.*, **12,** 682, December, 1951.

[2] Meagher, J. R., and H. J. Markley, "Practical Analysis of Ultra-high-frequency," RCA Service Company, 1943. This pamphlet gives a compact clear treatment of transmission lines and wave guides.

[3] McArthur, E. D., Disk-seal Tubes, *Electronics*, February, 1945, p. 98.

[4] Von Ohlson, L. H., The Small Signal Performance of the 416-B Planar Triode between 60 and 4,000 Mc, *Trans. IRE*, **ED-1** (4), 189, December, 1954.

up to 3,000 Mc good performance is also afforded by the cylindrical-structure "pencil" types of triodes.[1]

The short lengths of coaxial transmission lines and the large circular contacts required for connection to these u-h-f tubes give equipment of this type an appearance which has led to the term "plumbing" as the usual informal name. The characteristics and name continue to apply with preselectors and local oscillators of the transmission-line or wave-guide type up into the s-h-f range.

157. Butterfly Tuners.[2] Tuners of the butterfly type, such as shown in Fig. 37, are continuously adjustable over a wide frequency range, designs of moderate size achieving ratios as great as 5:1 in frequency and larger designs achieving greater ratios. These tuners have been made and used at frequencies as low as 40 Mc and as high as 1,100 Mc or more. The construction is somewhat similar in appearance to a variable capacitor, with the exceptions that (1) the circuit inductance is built in, consisting chiefly of the circular-strap portions of the stator plates, (2) both connections are made to the stator, whence (3) there are no sliding contacts. When the rotor is turned toward increased meshing, the frequency is lowered by (1) increased inductance, resulting from removal of the conducting rotor plates near the inductive straps, and (2) increased capacitance in each of the two series capacitors (one-terminal-to-rotor and rotor-to-other-terminal) constituting the capacitive element of the tuner. The rotor turns through an angle of 90 deg for its full frequency range.

Fig. 37. Butterfly tuner. P_1 and P_2 = points for electrical connections; R = rotor; S = stator.

One design of butterfly tuner covering the range from 220 to 1,100 Mc has the characteristics shown in Table 15. This design has a diameter of 2½ in. It has 0.04-cm clearance between stator and rotor plates as they mesh, so that high voltage and power cannot be handled. It has five rotor and six stator plates.

Butterfly tuners are suitable for use where wide frequency ranges must be handled. In design for such service, the size and number of plates are made as large as possible while still getting the necessary top frequency and freedom from spurious modes of operation. The necessary l.f. is then obtained by reducing the air-gap clearance as far as required.

Table 15. Characteristics of Butterfly Tuning Element Covering Range of 220 to 1,100 Mc

Frequency, Mc	220	1,100
Inductance, μh	0.011	0.004
Capacitance, μμf	48	5
Q	650	300
Impedance, ohms	9,800	8,600

158. Technique for Frequencies Where R-f Stage Cannot Be Used. At some point in the frequency band between 500 and 3,000 Mc, it is found that the use of an

[1] Rose, G. M., D. W. Power, and W. A. Harris, Pencil-type UHF Triodes, *RCA Rev.*, **10**, 321, September, 1949.

[2] Karplus, Eduard, Wide-range Tuned Circuits and Oscillators for High Frequencies, *Proc. IRE*, July, 1945, p. 426.

r-f stage in a receiver results in lower gain or lower signal-to-noise ratio than without the stage.[1] Klystrons will produce amplification at these frequencies and are useful as transmitting amplifiers, but they are too noisy for use in receivers.

The receiver for frequencies above this point, therefore, has as its first elements the antenna, a passive coupling circuit, a superheterodyne oscillator, and a silicon-crystal mixer for converting the signal to i.f. A simpler type, called a *crystal video receiver* and analogous to the t-r-f receiver formerly used for broadcast reception, consists of only an r-f coupling circuit, detector, and the modulation-frequency or "video" amplifier; there is no local oscillator, mixer, or i-f amplifier. Such crystal video receivers are used for radar beacon service and other cases where small lightweight equipment is required.[2]

The reflex klystron[3] is generally used as the oscillator for receivers in the s-h-f range. The likelihood of requiring a.f.c. is apparent when it is noted that a band width of 10 Mc at a carrier frequency of 10,000 Mc amounts to only 1 part in 1,000. A.f.c. is therefore generally provided, the control system operating by placing on the reflex electrode of the klystron a suitable voltage so as to keep the i.f. of the receiver at its proper value.

For wide-band reception, such as for narrow sharp pulses, in this frequency range, extensive design procedures from the television and radar[4] arts are available and are described in Secs. 194 to 196. Pulses as short as 0.006 μsec, corresponding to less than 10 ft of path length in the transmission medium, have been produced and utilized for checking the quality of wave guides at 9,000 Mc. The receiver for these pulses consisted of three cascaded traveling-wave stages as r-f amplifiers, the detector, a distributed type of video amplifier, and a special oscilloscope.[5]

Considerable interest has been shown in the traveling-wave tube described by Kompfner in England[6] and Pierce and Field[7] in the United States. An experimental tube of this type made at the Bell Laboratories furnishes substantial amplification over the very wide band width from 3,200 to 4,000 Mc. For transmitting purposes the tube will furnish substantial amounts of power, and its noise characteristics are satisfactory. To adapt it for receiver use extensive efforts have been directed toward the perfection of quiet electron guns so as to obtain lower noise factors.[8] The RCA type 6861 has a noise factor of only 6.5 db with a gain of 25 db for the entire band from 2,700 to 3,500 Mc.

Designs of traveling-wave tubes for 50,000 Mc have had gains of 19 db, and for 58,000 Mc gains of 9 db, with noise factors in the region of 25 to 30 db.[9]

[1] Mellen, G. L., *et al.*, UHF Long-range Communication Systems, *Proc. IRE*, **43**, 1269, October, 1955; see p. 1276 on noise factor with and without r-f stage. Reich, H. J., *et al.*, "Microwave Theory and Techniques," Van Nostrand, 1953.

[2] Henning, R. E., Microwave Video Detection Characteristics of Crystals, *Aeronautical Electronics Digest*, 1955, pp. 96–101. Moxon, L. A., "Recent Advances in Radio Receivers," Cambridge, 1949; see pp. 142–145 on crystal video receivers.

[3] Pierce, J. R., and W. G. Shepherd, Reflex Oscillators, *Bell System Tech. J.*, July, 1947, p. 460. Harrison, A. E., "Klystron Tubes," McGraw-Hill, 1947. McNally, J. O., and W. G. Shepherd, Reflex Oscillators for Radar Systems, *Proc. IRE*, December, 1947, p. 1424. Rabinowitz, S. J., Stabilization of Reflex Klystrons by High-Q External Cavities, *IRE Trans.*, **MTT-2** (3), 23, September, 1954.

[4] Schneider, Edwin G., Radar, *Proc. IRE*, August, 1946, p. 528; see section on receivers, pp. 551–556. Miller, S. E., Considerations in the Design of Centimeter-wave Radar Receivers, *Proc. IRE*, April, 1947, p. 340.

[5] Beck, A. C., Microwave Testing with Millimicrosecond Pulses, *Trans. IRE*, **MTT-2** (1), 93, April, 1954.

[6] Kompfner, R., The Traveling-wave Tube as Amplifier at Microwaves, *Proc. IRE*, February, 1947, p. 124.

[7] Pierce, J. R., and L. M. Field, Traveling-wave Tubes, *Proc. IRE*, February, 1947, p. 108. Pierce, J. R., Theory of the Beam-type Traveling-wave Tube, *Proc. IRE*, February, 1947, p. 111. See also Muller, M., Traveling-wave Amplifiers and Backward-wave Oscillators, *Proc. IRE*, **42**, 1651, November, 1954.

[8] Robinson, F. N. H., Microwave Shot Noise and Amplifiers, *Trans. IRE*, **ED-2**, 128, July, 1956. Pierce, J. R., "Electrons, Waves and Messages," Hanover House, 1956; see pp. 158–165 and 176–180. Mungall, A. G., Noise in Traveling-wave Tubes, *Trans. IRE*, **ED-2** (2), 12, April, 1955. Knechtli, R. C., and W. R. Beam, Performance and Design of Low-noise Guns for Traveling-wave Tubes, *RCA Rev.*, **17**, 410, September, 1956.

[9] Robertson, S. D., An Experimental Broad-band Helix Traveling-wave Amplifier for Millimeter Wavelengths, *Trans. IRE*, **MTT-2** (3), 48, September, 1954.

A combination of a traveling-wave tube and an oscilloscope, called a *wamoscope* (for wave-modulated oscilloscope), has been developed and affords an almost complete receiver in the single glass envelope. The Sylvania type 6762 is about 23 in. long and can be used at any frequency from 2,000 to 4,000 Mc. The output beam from the traveling-wave portion of the tube enters an electron-optical system which acts as a velocity-sorting detector; from here the beam goes on to the fluorescent screen as in the usual oscilloscope tube.

An important feature of microwave practice is the use of suitable ferrites in wave guides for such purposes as rotating the plane of polarization and thus acting as an attenuator.[1]

SUPERREGENERATIVE RECEIVERS

Superregeneration offers high gain, such as 100 db, in a single tube. Fundamentally a method of amplification, it is easily adapted to performing a-m detection also and has been very frequently used with this additional feature. On account of the high gain obtainable, it is possible to make a set of high sensitivity with only a few tubes. Superregeneration, described by Edwin H. Armstrong in 1922,[2] was little used except in amateur v-h-f receivers until the Second World War, when it was employed extensively in military equipment. Subsequently it was given further careful study.[3]

159. General Principles. The superregenerator tube is provided with ample feedback so as to be capable of oscillation at the desired r.f. It is also provided with a periodic quench voltage by which an oscillation which has built up is stopped for a time, after which the quench is removed and an oscillation builds up again. In this way the circuit is prevented from going into permanent oscillation; instead it engages in a periodic building up of transient oscillations whose average value depends upon the signal voltage existing in the circuit when each build-up starts. Thus amplification is produced. Superregeneration can therefore be considered as a sampling process by which samples of the signal are taken at quench-frequency intervals and much amplified. This process is sufficient to recreate the envelope if the highest modulation-frequency component of the signal is less than half the quench frequency.

In case no signal is present, fluctuation noise will start the oscillations, and this noise will therefore be amplified by the superregenerative action. However, impulse noise, which occurs at discrete instants, generally does not coincide with the sensitive periods of the superregeneration, so that most of the impulses are suppressed.

For purposes of analysis and design, the action of the superregenerative tube can be considered as presenting to the tuned circuit a negative resistance, or the equivalent negative conductance, during the build-up time. Conversely, during the quench interval the circuit has loss corresponding to a positive conductance. The result is that the conductance of the tuned circuit varies with time at the quench frequency. H-f oscillations build up during the period of negative conductance and decay during the following period of positive conductance. The conductance-time characteristic is one of the fundamental quantities determining the performance. Another important quantity is the total tuned-circuit capacitance.

The quench voltage can be provided by a separate oscillator, in which case the circuit is described as having "separate quench." Certain operating features are obtainable only in this way. Often, however, the superregenerative tube is provided with a suitable grid leak and capacitor so that its oscillations block periodically, thus ena-

[1] Healy, D. W., Jr., and R. A. Johnson, "Ferrite Applications: Electronic Properties of Ferrites and Their Application to Microwave Devices," Publication PB-122,890 of Office of Technical Services, Washington 25, D.C., 1956.

[2] Armstrong, Edwin H., Some Recent Developments in Regenerative Circuits, *Proc. IRE*, August, 1922, p. 244.

[3] Hazeltine, Alan, D. Richman, and B. D. Loughlin, Superregenerator Design, *Electronics*, September, 1948, p. 99. Bradley, William E., Superregenerative Detection, *Electronics*, September, 1948, p. 96. Wheeler, H. A., "Wheeler Monographs," Wheeler Laboratories, Great Neck, N.Y., 1953; see Nos. 3 and 7. Whitehead, J. R., "Super-regenerative Receivers," Cambridge, 1950. Mostafa, A. E-S., and M. El-Shishini, A New Method for Treating Electron Tubes When Used as Superregenerative Detectors, *Communication and Electronics*, No. 6, p. 207, May, 1953, and No. 7, p. 283, July, 1953.

bling the tube to serve as its own source of quench voltage. Such operation is described as "self-quenched."

There are two distinct modes[1] of superregenerative operation: the linear and logarithmic.

In the linear mode the quench frequency and other factors are so proportioned that the quench voltage acts to stop the building up of oscillations while they are still rising; *i.e.*, the oscillations do not reach the maximum amplitude at which the tube would operate if it were in a steady state of oscillation. The linear mode is given this name because the amplitude reached is proportional to the signal amplitude at the start of the build-up period; in this way there is a linear mathematical relation between output and input. Separate quench is required for the linear mode.

In the logarithmic mode the oscillations are allowed to build up to the maximum amplitude of which the tube is capable before the quench voltage is applied. Build-ups starting from various signal strengths are, however, not identical, because one started by a strong signal has a larger value throughout the entire build-up period and reaches the maximum value sooner than one starting from a weak signal. The modulation-frequency signal obtained upon detection is found to be proportional to the logarithm of the input signal amplitude, so that this type of operation is called the "logarithmic mode." This mode can be obtained with either separate or self-quench.

The three possible types of superregenerative circuits with respect to mode and quenching are shown in Table 16.

A logarithmic-mode superregenerator can easily be adapted to perform a-m detection in addition to superregenerative amplification, by providing an impedance in either the grid or plate circuit across which the m-f voltage can be developed.

In self-quenched operation the quenching action can occur only as a result of the oscillations building up to the maximum amplitude which is possible under the particular conditions. Such operation is therefore always of the logarithmic mode. With

Table 16. Main Classification of Superregenerative Circuits with Respect to Mode and Quenching

Quench	Mode	
	Linear	Logarithmic
Self-quench	Impossible	Often used. See Sec. 162.
Separate quench	Often used. See Sec. 160.	Can be used. See Sec. 161.

a strong signal the maximum amplitude is reached sooner, the quench voltage comes on sooner, and after the usual quiescent period, the next build-up of oscillations is sooner—all in comparison with the action for a weak signal. Therefore, the quench frequency varies in self-quenched operation, the value rising and falling with the amplitude of the signal.

If a superregenerative receiver oscillates at or near the signal frequency, it may be an objectionable source of radiation to other nearby receivers tuned to receive the same transmission. This radiation can be reduced by using an untuned shielded r-f stage ahead of the superregenerator.

The wave form of the quench voltage may be of several different shapes: sine-wave, rectangular, trapezoidal, saw-tooth, etc. The resulting curve of circuit conductance as a function of time will vary accordingly, and the performance may thus be considerably affected. As an example, selectivity is better for sine-wave quench than for rectangular-wave quench.

[1] Frink, F. W., The Basic Principles of Superregenerative Reception, *Proc. IRE*, January, 1938, p. 76.

Values of the quench frequency vary considerably according to the particular service, with a range of at least 15 to 2,000 kc.

160. The Linear Mode. Operation in the linear mode requires a separate quench. The gain in this case and the selectivity for either the linear or the logarithmic mode are obtainable from a knowledge of the circuit-conductance curve and the total circuit capacitance. The superregenerative gain is $4.35a/C$ db, where a is the area of the negative-conductance portion of the conductance-time curve (Fig. 38) and C is the total circuit capacitance.

Example: If a is 325×10^{-12} mho-sec and C is 20 $\mu\mu$f, the superregenerative gain is 70 db. This is to be added to the nonregenerative gain in the input coupling system and to the regenerative gain. Thus, if the nonregenerative gain is 20 db and the regenerative gain is 14 db, the total gain would come to an over-all figure of 104 db.

Fig. 38. Trapezoidal quench wave showing quantities which determine the sensitivity and selectivity of a superregenerator.

The band width for either linear or logarithmic operation, measured at 8.7 db (1 neper) down, is given by

$$\frac{1}{\pi}\sqrt{\frac{g'}{C}}$$

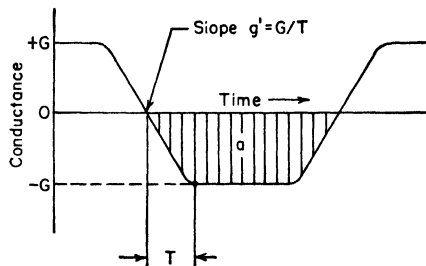
where g' is the rate (expressed as a positive quantity) at which the conductance is changing when the sign of the conductance changes from plus to minus, and C is the total capacitance of the circuit. This formula can be used with saw-tooth, trapezoidal, and sine-wave quenches, but not with square-wave quench; *i.e.*, the value of g' must not be too large. In particular, this formula is useful when the value of g' is less than $G^2/4C$, where G is the absolute value of the conductance at the end of the straight-line portion of the conductance-time curve, as shown in Fig. 38.

Example: If g' is 200 mhos per sec and C is 20 $\mu\mu$f, the 8.7-db band width is 1 Mc. The value of $G^2/4C$ is 405, so that the criterion for the applicability of the formula is satisfied.

Further information regarding band width is obtainable from the fact that the selectivity curve has the shape of a probability curve between the nose and a value of attenuation equal to $2.2GT/C$ db, where G and T (= time) have the values at the point indicated by a dot on the conductance curve of Fig. 38. Assuming, in the typical example, that the value of T is 0.9 μsec, the formula gives 18 db. In other cases this may amount to 40 to 80 db.

A probability curve has a shape corresponding to the selectivity of a large number of isolated single-tuned circuits in cascade. A superregenerative receiver can, therefore, have a selectivity curve with very steep sides. By comparison with the selectivity obtainable in conventional multicircuit receivers, the probability curve is somewhat broad near the resonant frequency; yet it has good skirt selectivity, which means good rejection of unwanted signals outside the pass band of the receiver. For the best selectivity a saw-tooth quench wave can be used with the gradual change from positive to negative conductance and the rapid change in the opposite direction. This decreases the value of g' and increases the value of T.

Automatic gain stabilization can be included in these receivers to maintain reception despite changes of antenna impedance, line voltage, tube aging, and the effects of high humidity. The system operates by detecting the quench-frequency pulsation of the r-f signal in the superregenerative circuit, amplifying it, rectifying it, and using the resulting d-c voltage to control the superregenerative tube. This maintains the quiescent noise output almost constant, in addition to keeping the general operation within good limits.

The sensitivity of a linear-mode superregenerator increases as the amplitude of the

quench voltage increases, but the quench amplitude must not be great enough to overload the tube. Sensitivity also increases as the frequency of the quench voltage is lowered. Selectivity increases, *i.e.*, band width decreases, for both lower frequency and lower amplitude of quench voltage. It is seen therefore that sensitivity and selectivity improve with lower values of the quench frequency. The requirements of sensitivity and selectivity with regard to amplitude are however opposed, low amplitude favoring selectivity and high amplitude favoring sensitivity.

161. Logarithmic Mode with Separate Quench. Theoretical and experimental studies of superregeneration of the logarithmic mode with separate quench have been made. There is considerable amplitude distortion with the logarithmic mode because a large increase of signal causes only a slight increase in the output due to the logarithmic output-input relationship. This is advantageous in reducing the effect of impulse noise, but produces considerable distortion in a-m reception. For headphone listening to speech signals the modulation under these conditions should therefore not exceed about 80 per cent.

162. Self-quenched Circuits in A-m Reception. All self-quenched circuits operate in the logarithmic mode, and the quench frequency is a function of the signal amplitude.

The self-quenched superregenerator has been extensively used in amateur practice, triodes such as the 6J5, 7A4, 9002, and 6C4 being employed with grid leaks of 2 to 10 megohms and grid capacitors of 50 to 100 $\mu\mu$f. The 6F4 is also very suitable. Quench frequencies have varied from 20 to 250 kc as the center value, with frequencies of 150 to 250 kc recommended for signals at 60 Mc and higher in the v-h-f range.

One refined amateur superregenerator[1] includes an untuned 6AK5 r-f stage and achieves a substantial reduction of radiation.

RECEIVERS FOR PULSE-TYPE COMMUNICATIONS

During the Second World War and the postwar period, there was an increasing realization that there are advantages in the use of pulses for general communication purposes,[2] such as multiplex telephony. Such a system utilizes the principle of sampling; *i.e.*, the signal wave to be transmitted is sampled at a frequency at least twice the highest component, so that the reproduced wave will be a satisfactory copy of the original.

163. Pulse Systems. Two of the several possible types of pulse systems are of importance: (1) pulse-time modulation (called PTM, or PPM for the alternative name pulse-position modulation) and (2) pulse-code modulation (PCM, also called pulse-count modulation). Both types use pulses of substantially rectangular shape and of constant amplitude. Since the sampling is at a frequency two or more times the highest signal component and since a number of harmonics are necessary to create the rectangular shape, the band width required by pulse systems is many times that needed in the conventional types of modulation. However, there are compensating advantages: (1) These systems are highly resistant to the effects of noise so that lower transmitting power can be used or a greater distance be obtained, and (2) in multiple operation where several messages are being sent simultaneously, nonlinear amplifiers can be used without danger of cross talk between the individual channels. It will be seen that these systems with pulses of constant amplitude have characteristics similar to telegraph practice where "regenerative" repeaters have long been used, the term "regenerative" referring to the restoring of the pulse shape in addition to the amplification in the repeater. On account of the simplicity of the wave shape, repeaters in pulse systems can produce such regeneration, so that the noise is substantially removed at each repeating point. In PTM and PCM systems the signal-restoring benefits

[1] Tilton, E. P., A Non-radiating Superregenerative Receiver for Two Meters, *QST*, February, 1946, p. 53.

[2] Pulse Communication, *J. IEE* (*London*), Part IIIA, No. 11, pp. 83–106, March and April, 1947. This gives one long paper and ten summaries of papers presented in March, 1947. Bandwidth vs. Noise in Communications Systems, *Electronics*, January, 1948, p. 72. How Rigid Is the Hartley Law? *Tele-Tech*, January, 1948, p. 52.

realized in telegraph practice can be obtained also in the much faster operations required for the transmission of signals for voice, facsimile, etc.

An important increase of the effective selectivity of a pulse receiver can be obtained by an invention[1] of the late R. B. J. Brunn. This utilizes the fact that a pulse of carrier frequency, according to Fourier analysis of the modulating wave, includes side frequencies near the carrier representing the pulse duration and the amplitude in the middle of the pulse and also includes well-spaced side frequencies representing the abrupt start and end of the pulse. If, therefore, such a pulse is detuned because it is at an interfering carrier frequency, the close-in side frequencies will be reduced, while the well-spaced ones on one side may be transmitted in fair amplitude. The effect of the selectivity is then to change the pulse shape from rectangular to a lower average level plus spikes, or ears, at the start and end. After detection, stabilization can be used to bring all positive peaks to the same level, and further processing will leave the undesired pulse represented only by spikes while the desired pulse has its normal shape. The use of a coincidence mixer then eliminates the spikes and delivers a useful output from the desired pulses.

164. Pulse-time Modulation. One method by which audio signals can be transmitted with pulses is by varying the position or time of successive pulses, each with respect to a reference pulse which occurs at a regular rate. Thus modulation is attained by displacing the signal pulse from its normal position or time by an amount proportional to the amplitude of the modulating signal. For example, if the pulse occurs later than normal, the audio amplitude may be positive; if the pulse occurs earlier, the amplitude will then be negative. It is possible to use a single reference pulse for numerous signal pulses, each signal pulse representing a different voice or other transmission so that multiplex operation is obtained.[2]

The effect of noise in a pulse-time system is limited to the extent to which it influences the time of occurrence of the signal pulse. As a result, the steeper the edges of the signal pulse, the less the influence of noise. Therefore, with increase of band width, which permits steeper slopes, noise has less effect on transmission. This system, therefore, obtains improved signal-to-noise ratio at the expense of greater band width.[3]

165. PTM System of 24 Channels. A 24-channel PTM system, developed by the Federal Telecommunications Laboratories,[4] required the transmission of modulating signal components up to 2.8 Mc. This system, suitable for use over a coaxial cable or over a radio link, utilizes pulses of 0.5 μsec duration, two of these with a 1.3-μsec separation constituting the marker pulse. Each signal pulse varies in position over a 2-μsec range in a 5-μsec time channel. Each of the signal channels is sampled 8,000 times per sec, corresponding to a 125-μsec interval. This 125 μsec is divided into 5-μsec time channels to accommodate the 24 multiplex channels and the marker channel. The frequency of pulses, including those for all channels, is therefore 8,000 × 25 = 200,000 cycles, or 200 kc, so that the band width of 2.8 Mc includes the first 14 harmonics and permits a fairly good reproduction of the rectangular pulse shape.

The Federal PTM system includes two Cyclophon tubes, one of which is used as the commutator in the modulator of the transmitter. The other is used as separator and demodulator in the receiver. This tube is a modified cathode-ray tube in which the beam is whirled around in a circular path where it goes through successive holes in an aperture plate and falls on individual "dynode" plates connected to the various channels. The system has been operated by radio[5] in the u-h-f range of 1,225 to 1,325 Mc. The receiver utilizes a 1N23B crystal mixer and a 2C43 oscillator. An

[1] Brunn, R. B. J., U.S. Patent 2665378, filed Mar. 26, 1949, issued Jan. 5, 1954.

[2] Deloraine, E. M., and E. Labin, Pulse Time Modulation, *Electronics*, January, 1945, p. 100; also *Elec. Commun.*, **22**, 91, 1944.

[3] Moskowitz, S., and D. D. Grieg, Noise-suppression Characteristics of Pulse-time Modulation, *Proc. IRE*, **36**, 446, April, 1948.

[4] Grieg, D. D., and A. M. Levine, Pulse-time-modulated Multiplex Radio Relay System—Terminal Equipment, *Elec. Commun.*, June, 1946, p. 159.

[5] Grieg, D. D., and H. Gallay, Pulse-time-modulated Multiplex Radio Relay System—Radio-frequency Equipment, *Elec. Commun.*, June, 1947, p. 141.

image rejection of 72 db and a noise factor of 14 db are obtained. The i.f. is 30 Mc. The band width at 1 db down is 8 Mc so that the 2.8-Mc maximum signal frequency, corresponding to 5.6 Mc for both side bands, is easily accommodated.

166. Pulse-code Modulation.[1] The method of sampling which characterizes all pulse-type communication systems is used in pulse-code modulation with the additional feature that, after a sample is taken, it is quantized; *i.e.*, a determination is made as to which one of a definite number of amplitude classes it falls into. The number of these classes is some power of 2, such as $2^4 = 16$, $2^5 = 32$, etc., the particular number depending on the design of the system.

However, for the purpose of explanation, assume that only four classes are provided, which is 2 raised to the second power. In this case the four available amplitude classes consist of two for positive values of the signal and two for negative values. The four classes are transmitted by two pulses, two being the exponent. The four classes consist of (1) both pulses present, (2) only the first pulse present, (3) only the second pulse present, and (4) both pulses absent.

It has been found by experiment that satisfactory commercial telephone quality can be obtained by the use of a five-pulse system, which gives 2^5 or 32 amplitude classes. Of these, 16 are positive and 16 negative. Since the range of volume to be accommodated is about 60 db, this corresponds to approximately 2 db per amplitude class. Tests have shown this degree of quantizing to give a 33-db value of the signal-to-noise ratio; *i.e.*, some "quantizing noise" is present and prevents the signal-to-noise ratio from having a larger value.

In PCM the particular signal amplitude at a given moment is represented by the presence or absence of the individual pulses in a group of pulses. As an example, in a five-pulse system, if the first, third, and fifth pulses are present and the second and fourth pulses are absent, a particular amplitude value is indicated.

PCM is therefore characterized by the fact that all pulses transmitted have equal amplitude and equal duration; also (with the exception that certain pulses are omitted) they occur at equal intervals. Under these circumstances, noise is ineffective unless it is sufficiently strong to cause a pulse to appear where none was transmitted, or vice versa. This system is therefore highly resistant to noise and, under typical operating conditions, received signal strengths can be weaker than with other systems. At each repeater the restoring type of amplifier sends on the signal entirely free of noise; *i.e.*, noise is noncumulative as long as the incoming signal-to-noise ratio is above a modest threshold value.

In a typical system using five pulses or 32 amplitude values, the band width is somewhat less than five times that required for a pulse-amplitude-modulated system. This is the cost in band width to obtain the very large increase in ability to withstand noise.

On account of the complexity of the PCM equipment required to code the amplitudes at the transmitter and decode them at the receiver, the system is of special interest in connection with multiplex transmission, a time-division system being used in which the five pulses of one channel are transmitted, then the five pulses of the next channel, etc., the entire system operating with short pulses and short spacings so as to return to the first channel after an interval of 1/8,000 sec or 125 μsec. In a system having eight channels, this means that pulses are transmitted at the rate of $5 \times 8{,}000 \times 8$ or 320,000 per sec.

MOBILE RECEIVERS

167. General Technical and Regulatory Aspects.[2] Communication requirements of mobile services such as police, taxis, fire departments, public utilities, forestry

[1] Grieg, D. D., Pulse Count Modulation System, *Tele-Tech*, September, 1947, p. 48; material also given under title Pulse-count Modulation, *Elec. Commun.*, September, 1947, p. 287. Black, H. S., and J. O. Edson, PCM Equipment, *Elec. Eng.*, November, 1947, p. 1123. Clavier, A. G., P. F. Panter, and D. D. Grieg, PCM Distortion Analysis, *Elec. Eng.*, November, 1947, p. 1110. Coded Pulse Modulation Minimizes Noise, *Electronics*, December, 1947, p. 126. Pulse Code Modulation, *Electronics*, August, 1947, p. 136. Meacham, L. A., and E. Peterson, An Experimental Pulse Code Modulation System of Toll Quality, *Bell System Tech. J.*, January, 1948, p. 1.

[2] Macdonald, A. A., Comparison of Split Channel FM and Single Sideband for Land Mobile Services.

services, etc., differ among themselves, but in general the receivers follow conventional engineering practice with certain special features, such as (1) fixed-frequency operation, (2) location of the r.f. generally in the v-h-f range, (3) squelch (see Sec. 93) to keep the loud-speaker quiet between messages, and (4) convenient arrangements for changing between reception and transmission. Mobile radiotelephone apparatus often has facilities for the selective calling of particular cars and for tying into the general wire telephone system.

The regulations of the FCC governing these services in 1957 permit amplitude modulation with a total bandwidth of 8 kc for the two side bands and a 70 per cent normal modulation, or frequency modulation with a total band width of 40 kc and not over ±15 kc deviation. (However, frequency modulation is not permitted on carrier frequencies below 30 Mc.) With either type of modulation, the maximum audio frequency permitted to be transmitted is 3,000 cps.

Various ranges of carrier frequencies are allocated to the different mobile services by the FCC, but in general all these bands are severely congested. To provide additional channels in the region of 160 Mc, the Commission has adopted the rule that after October, 1958, all equipment produced must be capable of narrow-band operation, *i.e.*, operation with 15- or 30-kc channels instead of 60-kc channels as previously.[1]

Standard test procedures for mobile receivers differ substantially from those applicable to broadcast receivers and require careful attention by the designer of mobile types.[2] For example, the standard signal for testing the sensitivity of f-m receivers has a modulation of 70 per cent of full system deviation and a modulation frequency of 1,000 cps. Selectivity is measured with a two-signal method. Intermodulation is determined in terms of effect on the signal-to-noise ratio, and the measurement requires three signal generators.

Substantial amateur experience in mobile communications, including reception, suppression of automobile interference, and other phases, is available in the literature.[3]

168. Police Service. The usual police service installed in recent years operates with phase modulation, equivalent to f.m. with a preemphasis of 6 db per octave. Two-frequency operation is the general rule, with one frequency for headquarters-to-cars transmission and the other for cars-to-headquarters transmission. Car-to-car feature is then obtainable by having a double-frequency transmitter or a double-frequency receiver in each car. In extensive systems, several receiving stations are necessary, the signal picked up by the various receiving stations being relayed to headquarters by wire line or over radio relay links using other frequencies. This constitutes a special-type large-scale diversity receiving system. In state-wide police systems, several transmitting stations are also usually required.

To avoid interference between police systems in various cities, the power transmitted by the headquarters equipment is generally limited, so that, with small antennas on the cars, high sensitivity in the receiver is required.

A vertical antenna is generally used, mounted either at the rear of the car or in the center of the top. The latter position has the advantage of better omnidirectional

---

Trans. IRE, **PGVC-7**, 46, December, 1956. Bullington, K., Frequency Economy in Mobile Radio Bands, *Bell System Tech. J.*, **32**, 42, January, 1953. Young, W. R., Jr., Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc, *Bell System Tech. J.*, **31**, 1068, November, 1952. Robbins, M. A., and G. Ayer, A High Performance Mobile Unit for 450 Megacycles, *Trans. IRE*, **PGVC-6**, 43, July, 1956. Manke, A. G., Crystal Oscillators in Communication Receivers, *Trans. IRE*, **PGVC-7**, 10, December, 1956.

[1] Symposium of 14 papers on spectrum conservation in the mobile services, *Trans. IRE*, **PGVC-3**, 1, June, 1953.

[2] "Minimum Standards for Land-mobile Communication FM or PM Receivers," EIA Standard TR-119-A, June, 1955. "Standards on Vehicular Communications: Methods of Testing, 1949," IRE Standard 49 IRE 16.S 1, *Proc. IRE*, **37**, 1372, December, 1949; also designated ASA 16.18-1951.

[3] "The Mobile Manual for Radio Amateurs," American Radio Relay League, West Hartford, Conn., 1955.

characteristics.[1] Mounting at the back has the disadvantage that the antenna is most sensitive toward the rear.

169. Mobile Radiotelephone. Receivers used in the mobile telephone service are very similar to police equipment except that a handset and a telephone bell are generally provided; also selective calling is usually employed so that the bell rings only in the desired vehicle. Connections may be made to the general wire telephone network of the country, and use of the equipment on the vehicle is practically like the use of an ordinary phone with the exception that a press-to-talk button must be held down while speaking and released while listening. A large number of cars have the same pair of transmitting and receiving frequencies. On account of this "party line" type of operation, it is necessary for the telephone user in the car to listen to see if the facilities are in use before making a call.

Frequency assignments in the band from 30 to 44 Mc have been made for highway mobile telephone, and assignments in the band from 152 to 162 Mc for urban mobile telephone. One frequency for station to vehicle and another for vehicle to station are assigned and together are said to constitute a "channel" for this service. The method of transmission is phase modulation.

The requirements for marine radiotelephony, such as in coastal waters, are much like those in land mobile service with the exception that greater distances may require more powerful transmitters.[2]

170. Selective Ringing. For the purpose of ringing only the desired phone in the mobile service, two types of selective equipment are available. In one type each digit of the called number is represented by a group of the proper number of pulses, and the receiver can be said to operate by counting the pulses in the various groups. In the other type each digit is represented by a particular frequency, or "tone," or a combination of tones. In this case the receiver has filters for identifying the transmitted digits. When the particular installation is being called, the presence of the right digits is recognized by the equipment and the bell is automatically rung, or in industrial or police systems, a loud-speaker may be actuated, a light turned on, or a horn sounded.

An EIA standard[3] provides a definition of the sensitivity of a selective signaling system as the lowest carrier level (determined as in a normal receiver-sensitivity measurement) at which positive operation of the selector is obtained. The corresponding signal-to-noise ratio may then be obtained, keeping the same carrier level, by the following steps: (1) remove the calling modulation at the signal generator; (2) measure the noise output from the receiver; (3) apply 1,000-cps 100 per cent modulation and measure the receiver output; (4) compute the ratio of the outputs in steps 3 and 2, which is the measured signal-to-noise ratio. The result, to comply with the standard, must not be greater than 10 db.

The sensitivity of the selector itself, as distinguished from the entire receiver, is defined as the minimum power input of calling tone which affords positive operation. For this test the selector is connected directly to the tone generator through an impedance-matching element.

Regarding temperature and humidity exposure, the EIA standard requires that neither of the following conditions shall cause harm to the equipment or failure of operation: (1) temperatures from −30 to +60°C at 50 per cent relative humidity, and (2) humidity of 90 per cent and temperature of 50°C.

RAILROAD RECEIVERS

Radio equipment is used in railroad service in the following ways: (1) two-way radiotelephone for dispatching use, especially in freight yards and to a growing extent in main-line operation,[4] (2) receiver operation for the reception of broadcast programs

[1] Landis, C. G., and S. K. Waldorf, Radiation Pattern Study Shows Best Antenna Location Is "Center" on Car Roof, *Elec. World*, **139**, 66, May 4, 1953.

[2] Ornstein, W., Duplex and Multi-channel Mobile Equipment, *Trans. IRE*, **PGVC-4**, 22, June, 1954.

[3] "Minimum Standards for Land-mobile Selective Signalling Equipment," EIA Standard TR-120, May, 1950.

[4] Freedman, Samuel, "Two-way Radio," Ziff-Davis Publishing Co., 1946; note Chap. 9, Induction Radio and Guided Carrier Systems, pp. 143–173, and Chap. 12, Two-way Radio for Railroads, pp. 229–318.

for the entertainment of passengers, and (3) passenger telephone service enabling calls to be made to and from any phone in the regular land telephone network.

171. Dispatching Service.[1] Frequency assignments in the region from 158 to 162 Mc have been made for the operation of railroad dispatching telephone service. Vertical polarization and phase modulation are used.

In main-line freight service, equipment can be provided in both engine and caboose, affording time-saving communication between the two ends of the train. As an example of main-line operation, a 160-Mc 884-mile installation on the Erie Railroad used 50 base stations, or an average of 18 miles between base stations, the spacing being closer in hilly country. Towers 63 ft high with transmitters of 15 to 20 watts gave at least 5 μv at the receivers.

Flexible mounting of radio equipment is needed in railroad installations, whereas in automotive service it is generally not provided. Larger audio outputs, in the range from 4 to 10 watts, are needed in railroad service because of the high ambient noise.

The v-h-f dispatching service is often called "space radio" to distinguish it from systems having an inductive characteristic, which are taken up in the next paragraph.

172. L-f Inductive and Other Dispatching Services. Guided radio (*i.e.*, wired wireless) systems utilizing the rails and especially the normal wayside telegraph and telephone wires to propagate the waves offer the advantage of greater distance than obtainable with v-h-f operation. Distances as great as 50 to 100 miles between a wayside station and trains are typical.

These systems generally operate on frequencies between 50 and 200 kc and use frequency or phase modulation with a value of approximately unity for the modulation index.

In extended railroad yards an inductive system can be used for the remote control of an additional v-h-f transmitter-receiver. In one design[2] a carrier of 189 kc was used with a deviation of ±4 kc. The distance between the main and the satellite stations was 12 miles. The receiver was a t-r-f type requiring 10 μv to saturate, under which condition it delivered an output of 5 watts; the band width at 50 db was 20 kc and at 70 db was 26 kc.

A railroad dispatching system operating at 2,660 Mc[3] has been reported to give good results. The receiver in this equipment employs a crystal mixer followed by a 7-Mc i-f amplifier having four regular stages, followed by two limiter stages. The i-f band width at 6 db is 300 kc. A special vertical antenna array of biconical parabolic reflectors was designed for this service.

At the opposite extreme in the frequency spectrum is the unit train communication system of the Union Switch and Signal Company.[4] This is an a-m system in which the transmitted signal is a single upper side band of a suppressed carrier such as 5,700 cps. This signal at very low voltage and high current is fed conductively into the rails and picked up inductively from the rails. If paralleling telegraph or phone wires on the right of way are present, distances up to 10 miles can be obtained between trains and distances up to 100 miles between fixed stations and trains. The receiver in this system has two stages which amplify the received single side band and have a.v.c. A push-pull mixer, a peak limiter, and audio amplifiers are also provided.

173. Broadcast Reception on Passenger Trains. Conventional home receivers have long been used in limited numbers on railroads for the reception of programs from standard m-f broadcast stations, the special provisions of such installations generally being limited to suitable antenna and power-supply arrangements. Recently special designs for this service have been made with very good a.v.c., low spurious responses, auxiliary tape sources of music, etc.[5]

[1] Sands, L. G., VHF Radio Equipment Speeds Up Railroad Operation, *Tele-Tech*, May, 1947, p. 380. Peterson, Joel, Railroads Plan Greater Use of Radio for Communications, *Tele-Tech*, January, 1947, p. 78.

[2] Halstead, W. S., Railroad F-M Satellite System, *Communications*, May, 1946, p. 17. Nygren, A. C., and W. G. Clinton, 161-Mc Satellite System for Railyards, *FM & Television*, April, 1946, p. 38.

[3] Dahl, E. A., 2660-Mc Train Communication System, *Electronics*, January, 1946, p. 118.

[4] Grondahl, L. O., and P. M. Bossart, Train Communication, *Elec. Eng., Trans. Sec.*, July, 1943, p. 493.

[5] Klippel, K. L., and E. A. Dahl, Railway Entertainment System, *Electronics*, May, 1947, p. 118. Curtis, John A., Passenger Entertainment System for Railroad Use, *Tele-Tech*, July, 1947, p. 34.

174. Telephone Service on Trains. Telephone calls from passengers on railroad trains are handled by the urban and highway mobile facilities, the equipment being similar except for refinements which eliminate the necessity for the railroad user to operate a push-to-talk switch.[1]

MARINE, AMATEUR, AND POINT-TO-POINT RECEIVERS

Receivers designed for use in marine, amateur, and point-to-point service are often called "communication receivers"[2] to distinguish them from broadcast receivers. They differ markedly from broadcast sets in various respects. The general quality of construction is superior, particularly with regard to the use of longer-lived components and to the frequent provision of increased selectivity for reception in the presence of heavy interference. An outdoor antenna rather than a loop antenna is used. Various special features and numerous controls are provided. The housing is generally a functional metal cabinet. In point-to-point receivers, special features and individual design for the particular service may be extensive.

175. Marine Receivers.[3] The severe climatic conditions encountered in tropical waters make it necessary for marine receivers to utilize parts treated with special impregnating agents. In addition, a high degree of general reliability is required, as well as accessibility in case repairs are needed.

Legal requirements applicable to shipboard radio equipment are given in an FCC publication.[4] Large vessels must have separate main and emergency radiotelegraph receivers with separate antennas. Each of these receivers must be capable of receiving c-w and i-c-w telegraph from 100 to 200 kc and from 485 to 515 kc for signal inputs of 100 μv or more. For radiotelephone reception a receiver is required for 1,600 to 3,500 kc with a sensitivity of 50 μv in terms of the more difficult of the two requirements of (1) 0.05-watt output, or (2) 6-db signal-to-noise ratio. Other requirements apply to vessels subject to the Great Lakes Agreement and to equipment for lifeboat use.

Complete assemblies of transmitters and receivers of suitable types are available for shipboard installations.[5]

Another type of marine equipment for ocean-going service is available to provide two-way telephone service at frequencies from 2 to 12 Mc, and especially for the range from 2 to 3 Mc. Except for the use of a.m. and the larger and heavier construction, equipment of this type is similar to that provided for the mobile telephone service. Equipments in this service often include preset tuning arrangements providing a choice of six or ten frequencies with facilities to change the tuning of both transmitter and receiver simultaneously. A push-to-talk button is provided, or a voice-operated relay may perform this function automatically. Selective ringing equipment is available as an additional feature.

There is a trend toward the use of the suppressed-carrier single-side-band method in marine radiotelephone service[6] (see Sec. 152). The intelligibility is satisfactory if the restored carrier is within about 150 cps of the correct value, but for naturalness in the reproduction the difference should be less than about 40 cycles. Receivers for single-side-band service can employ a mechanical i-f coupling system (Sec. 74), which offers adequate band width near resonance plus extreme skirt selectivity.

[1] Monk, N., and S. B. Wright, Technical Aspects of Experimental Public Telephone Service on Railroad Trains, *Proc. IRE*, **36**, 1146, September, 1948.

[2] Grisdale, G. L., and R. B. Armstrong, Tendencies in the Design of the Communication Type of Receiver, *J. IEE (London)*, Part III, September, 1946, p. 365.

[3] Nilson, A. R., and J. L. Hornung, "Practical Radio Communication," 2d ed., McGraw-Hill, 1943. See Chap. 14, pp. 722-753.

[4] Federal Communications Commission, "Rules and Regulations, Part 8—Stations in the Maritime Services," sold by Government Printing Office, Washington 25, D.C.

[5] Standardized Marine Radio Unit, *Electronics*, January, 1942, p. 36. Lewis, E. F., Packaged Marine Radio, *Electronic Inds.*, October, 1943, p. 88. Byrnes, I. F., Recent Maritime Radio and Radar Developments, *RCA Rev.*, **14**, 305, September, 1953.

[6] Pappenfus, E. W., Single Sideband Techniques for Marine Communications, *Trans. IRE*, **CS-3** (1), 50, March, 1955. Laport, E. A., and K. L. Neumann, A New Low-power Single-sideband Communication System, *RCA Rev.*, **16**, 635, December, 1955.

Marine service does not normally afford sufficient room for useful employment of space diversity reception but can be benefited by the use of transmitting diversity,[1] *i.e.*, the use of two shore transmitters simultaneously sending the same message to the one shipboard antenna and receiver, as discussed in Sec. 193.

Marine receivers arranged for auto-alarm service at 500 kc are treated in Sec. 187. Receivers for direction finding in marine and other services are discussed in Secs. 184 to 186.

176. Amateur Receivers.[2] The manufactured sets offered for amateur service average about 12 tubes and provide various special features. A tuned r-f stage is generally included, to give improved image and i-f rejections and signal-to-noise ratio. A band-spread knob and dial are provided and permit band-spread operation at any point in the frequency range of the set, or in some designs in the amateur bands only. A switch to disable the a.v.c. and one to start a beat-frequency oscillator adapt the set for the reception of c-w telegraph signals.

The crowded conditions on the amateur bands make high selectivity a prime requirement.[3] For this reason, the amateur type of receiver includes in the i-f amplifier a filter having a quartz crystal as one element. The usual i.f. for an amateur receiver is 455 kc, so that this is the natural frequency of the crystal. The very sharp mechanical resonance of this crystal permits band widths as small as 200 cycles to be obtained, thus permitting the reception of telegraph signals through very severe interference. This very sharp selectivity has led to the name "single-signal" receiver. A selectivity control is provided to permit various greater band widths so that phone messages may be received with the best quality permitted by the prevailing conditions.

A noise limiter is generally found in the amateur receiver and is useful for reducing ignition interference.

177. Point-to-point Receivers. The particular requirements often existing in point-to-point service have led to many special receiver designs, some of which are very elaborate. The use of diversity antenna systems is found in this service (see Sec. 193).[4]

An outstanding case of point-to-point receiver design is the transoceanic radiotelephone equipment used by the Bell System.[5] The receiver in this equipment operates on any frequency from 4 to 22 Mc and provides as many as four one-way voice channels in a radio band having a total width of 12 kc. Each channel is of commercial telephone quality with a width of about 3,000 cycles. An elaborate form of single-side-band operation (Sec. 152) is used.

These channels are obtained as follows: (1) a side band somewhat removed from the carrier on one side furnishes one channel; (2) a similar displaced side band on the other side of the carrier furnishes the second channel; and (3) a third channel is "straddled" across the carrier, or the third and fourth channels are placed near the carrier on the two sides.

All channels of such transoceanic telephone services normally utilize the standard commercial privacy equipment (see also Sec. 200), which divides the voice range of 250 to

[1] Hansell, G. E., Transmitter Space Diversity as Applied to Shipboard Reception, *Trans. IRE*, **CS-3** (1), 44, March, 1955.

[2] "The Radio Amateur's Handbook," 35th ed., American Radio Relay League, West Hartford, Conn., 1958; see pp. 89–146 and 397–417. "The Radio Handbook," 14th ed., Editors and Engineers, Ltd., Summerland, Calif., 1956; see pp. 159–188 and 476–501.

[3] Villard, O. G., Jr., and D. K. Weaver, Jr., The "Selectoject," *QST*, **33** (11), 11, 100, November, 1949, describes an audio phase-balancing rejection or reinforcing circuit.

[4] Dietsch, C. G., The Tangier Radio Relay System, *RCA Rev.*, **14**, 557, December, 1953. See also same author's paper, The Tangier Radio Relay System of RCA Communications, Inc., *Trans. IRE*, **CS-2** (1), 65, January, 1954.

[5] In this connection, see A. H. Mumford, Long-distance Point-to-point Communication, *J. IEE (London)*, Part III A, 1947, pp. 23–42; W. J. Bray, H. G. Lillicrap, and W. R. H. Lowry, The Design of Transmitter Drives and Receivers for Single-sideband Systems, *J. IEE (London)*, Part IIIA, 1947, pp. 298–312; van der Wyck, C. F. T., Modern Single-sideband Equipment of the Netherlands Postals Telephone and Telegraph, *Proc. IRE*, **36**, 970, August, 1948; Schlaack, N. F., Development of the LD Radio System, *Trans. IRE*, **CS-2** (1), 29, January, 1954; Stiles, K. P., Overseas Radiotelephone Services of A. T. and T. Co., *Trans. IRE*, **CS-2** (1), 39, January, 1954.

3,000 cps into five equal 550-cps portions and interchanges these, with or without inversion, making a change in the arrangements every 20 sec.

To receive and separate a transmission of this kind is the function of a receiver having various special features. The design is a double superheterodyne with a first i.f. of 2,800 or 2,900 kc and a second i.f. of 100 kc. The transmission employs carrier of reduced amplitude. The receiver can therefore be operated to produce and use reconditioned carrier obtained from the incoming signal, or it can produce and use a standard 100-kc frequency as a resupplied carrier. Crystal filters operating in the neighborhood of 100 kc provide sufficient selectivity for separation of the carrier and the various desired side bands. The receiver includes a mechanical type of a.f.c. which operates on the first heterodyne oscillator so as to maintain the frequency of the 100-kc carrier in the second i-f amplifier.

Studies of long-distance ionospheric (25 to 60 Mc) communication by Murray G. Crosby[1] led to the conclusion that "the trend in improving the difficult transmission conditions of long-range communication has been toward the angle-modulated frequency shift system in the case of telegraphy, towards single-sideband operation in the case of a band-type of modulation, and towards improved applications of exalted-carrier and single-sideband diversity receiving systems."

One of the severe requirements applying to both transmitter and receiver in multiplex radiotelephone systems is to hold intermodulation to a sufficiently low level so that the signal in one channel will not cause excessive noise in the other channels. A 150-Mc system using f.m. with ± 15 kc maximum deviation, providing the equivalent of four phone channels, has an intermodulation between the two test tones which is 50 db down, a value considerably better than that of f-m mobile service or single-channel point-to-point phone service.[2]

One of the features needed for receivers in radio extensions of land telephone service is CODAN (carrier-operated device anti-noise).[3] This is a squelch system (Sec. 93) serving to disable the receiver when the person at the other end of the connection is not talking.

Another special feature often forming a part of the receiver is the expanding portion of a compandor (Sec. 123) to restore the volume range after the compression performed at the transmitter. Available companding equipment gives an average noise reduction of 22 db.[4]

Other devices used in such radiotelephone service and having an influence on receiver design are the VOGAD (voice-operated gain-adjusting device), which can be called an "audio automatic volume control," and the VODAS (voice-operated device anti-singing), which is a system of voice-operated relays to disable the unused direction of transmission at each instant so as to prevent oscillation or objectionable echoes.

An outstanding type of point-to-point reception is in *scatter-propagation* systems,[5] which are characterized by: (1) use of frequencies in the v-h-f, u-h-f, and s-h-f bands (which are normally limited to horizon distances); (2) operation, nevertheless, at distances very much beyond horizon values; (3) use of high transmitting power; (4) use of large directive antennas at both transmitter and receiver; and (5) use of more than one receiving antenna in a receiver-diversity system. Under these circumstances the receiver should be designed for low noise factor and good diversity operation. At frequencies of 100, 200, and 400 Mc noise factors of 3, 3.5, and 4.5 db, respectively, are obtainable, and at frequencies above 900 or 1,000 Mc, where an r-f stage is no benefit, a value of about 8 db is obtainable.

178. U-h-f and S-h-f Relay Receivers. A relay system operating at 2,000 Mc has been installed by General Electric between New York and Schenectady.[6] This

[1] Crosby, M. G., Long-range Communication Trends, *Trans. IRE*, **CS-1** (1), 41, July, 1953.

[2] McSweeney, R., Very-high-frequency Radio Link between Puerto Rico and the Virgin Islands, *Communication and Electronics* (AIEE), No. 22, pp. 781–785, January, 1956.

[3] Rudd, J. B., A CODAN for A-M Receivers, *Trans. IRE*, **CS-2**, 45, January, 1954.

[4] Harp, M. C., M. H. Kebby, and E. J. Rüdisuhle, Application of Compandors to FM Radio Systems with Frequency Division Multiplexing, *Trans. IRE*, **CS-2** (2), 36, July, 1954. See also Chap. 28 of this handbook.

[5] Scatter-propagation issue of *Proc. IRE*, **43** (10), 1171, October, 1955, gives 34 papers on the subject. Kamen and Doundoulakis, "Scatter Propagation," Howard W. Sams & Co., 1956.

[6] Deerhake, F. M., 2000-Mc Television Program Chain, *Electronics*, February, 1948, p. 94.

system operates with f.m. and a maximum deviation of ±7 Mc. The receiver resembles those in radar practice and has a 1N21B crystal mixer, an SRL-7A klystron superheterodyne oscillator, and an i-f amplifier operating at 100 Mc. The i-f amplifier includes three triplets, each consisting of three stages. Each triplet is stagger-tuned at 92, 100, and 108 Mc. In this way a band width of 16 Mc is obtained. These three triplets are followed by a two-stage limiter, which is followed by a discriminator for detection of the video signals, the purpose of the relay system being to transmit television programs.

A 4,000-Mc relay system for television and multiplex telephone service has been designed by the Bell System and widely installed. The receivers in this system utilize a crystal for conversion to an i.f. of 70 Mc. No r-f amplification is used in the receivers. The transmitters use the planar-electrode 416-B triode, which in other applications (Sec. 156) has been found suitable for small-signal amplification at frequencies from 60 to 4,000 Mc.

Point-to-point equipment for television pickup service at 7,000 Mc has been designed by RCA.[1] This operates with f.m. The receiver is divided into two parts, the first of which is mounted directly behind the parabolic "dish" which constitutes the receiving antenna. This part of the receiver includes a crystal mixer, a 2K26 klystron oscillator, and four i-f stages operating at 120 Mc. The band width is 20 Mc. A.f.c. with a d-c amplifier is provided, acting on the reflex electrode of the klystron. The second unit of the receiver furnishes additional i-f amplification, as well as discriminators for the signal circuit and for the a-f-c operation. The i.f. corresponding to the peaks of the television synchronizing signal is held by means of the a.f.c. to an accuracy of ±1 Mc regardless of the picture content.

MILITARY AND NAVAL RECEIVERS

Special attention to reliability despite severe conditions of temperature, humidity, mechanical shock, and vibration is necessary in the design of military and naval radio equipment. In addition, in case the apparatus is for use in aircraft, it must operate at greatly reduced air pressure or be housed in an airtight case raising problems of heat dissipation. Numerous specifications have been prepared giving details of these various requirements, and compliance with such specifications is required according to the contracts under which equipment is purchased by the Government from manufacturers. Special attention is necessary on the part of the design engineer to ascertain definitely what specifications must be observed and to work out satisfactory designs.

A systematic method of assigning model designations is used in the United States defense services.[2] For example, in the *AN/APR*-3 the *AN* indicates that this is a military equipment, the *A* following the slant line indicates that it is for aircraft installation, the *P* indicates radar (or "pulse") equipment, the *R* indicates the purpose as reception, and the 3 distinguishes from preceding and later assignments of the same letters.

179. Typical Military Requirements. Selected provisions from various specifications will serve to indicate some of the special features required in radio equipment for the defense services. The precise requirements in each case depend on the specifications listed in the contract applicable to the particular equipment, the choice of the particular specifications being made by the Government authorities on the basis of the character and purpose of the equipment and the state of military and radio knowledge at the time.

Temperature. A typical requirement is that operation must be satisfactory for extended periods at any point from −55°C (−67°F) to +55°C (+131°F). Operation at higher temperatures may be required in some cases. Resistance to high humidity is generally required, such as exposure to 100 per cent humidity at a temperature of 50°C (122°F).

[1] Poch, W. C., and J. P. Taylor, Microwave Equipment for Television Relay Service, *Broadcast News*, October, 1946, p. 20.

[2] "Reference Data for Radio Engineers," International Telephone and Telegraph Corp., New York, 1956; see pp. 957–963.

Pressure. For air-borne equipment, operation at a pressure of 3.4 in. Hg, corresponding to an altitude of 50,000 ft, is a representative requirement.

Vibration. Requirements may include the range from 10 to 55 cps with the equipment being tested at a total excursion of 0.06 in. The design of flexible mountings for resisting vibration and shock is quite important.[1]

Shock. Resistance to shock is necessary so that the equipment shall not be readily disabled by the rough landing of aircraft, enemy bombs bursting on the decks of vessels, or other causes. Shock corresponding to an acceleration of 25 g is sometimes stated as a value which must be successfully withstood. Another specification, in dealing with equipment weighing over 250 lb, requires that it be mounted on a table weighing about 2 tons, and that the table then be struck underneath with a 3,000-lb hammer moving at a velocity corresponding to a fall through a height of 9 to 66 in., the value of this equivalent fall varying with the weight and class of the equipment.

Components. Inductors and transformers having laminated-iron cores are subject to severe requirements and must pass tests for temperature rise of the windings in operation, as well as exposure to extreme external temperatures and immersion in successive baths of hot and cold water. The potting compound must not crack at −55°C (−67°F) nor flow at +85°C (+185°F).

Variable capacitors must have a nominal clearance between fixed and moving plates of not less than 0.012 in. and an actual clearance not less than 0.008 in.

Electrolytic capacitors are normally usable only if approved by the Government after a showing by the contractor. In case their use is approved, the plug-in type is preferred but must be provided with means for firmly holding the capacitor in its socket.

Resistance to corrosion must include the ability to withstand exposure to a salt sea atmosphere. In the case of steel, phosphate treatment is generally acceptable. Another acceptable finish on steel is a cadmium plating of 0.0005 in. thickness followed by either a dichromate treatment or lacquering. Aluminum can be treated by (1) anodizing, (2) a caustic dip and lacquer, or (3) sandblast and lacquer.

Major improvements have been made in components[2] to meet the military needs for smaller size, improved reliability, ability to withstand higher temperatures,[3] and other properties.

The transistor offers various advantages for military equipment. A study of its application in air-borne apparatus led to the recommendation of 18 volts as standard d-c power supply for Air Force transistor equipments.[4]

Fungus. The war in the Pacific demonstrated that resistance to the growth of fungus[5] is one of the most important requirements for military equipment to be operated in the tropics. It is necessary to avoid the use of fungus-nutrient material as far as possible and, where this is impossible, to treat with a fungicide. The thickness of the fungicide spray after drying should be 0.002 in.

Radiation. Receiver radiation must be less than 400 $\mu\mu$w according to the usual naval requirement.[6]

Soldering. The soldering of wires must normally follow a secure mechanical attachment, so that the soldering is not depended on for mechanical strength. The use of excessive solder is not acceptable. Not over three wires may be attached to one lug, but multiple lugs may be used to whatever extent necessary.

Preferred Circuits. A study of many circuits for the most frequent functions was made by the National Bureau of Standards and led to the recommendation of 32 as preferred types for defense equipment.[7] These include power supplies, amplifiers, voltage regulators, and other types.

[1] Mercurio, S. P., Jr., and M. Belby, Designing Airborne Equipment for Vibration, *Electronics*, **30**, 166, January, 1957. Crede, C. E., Shock Testing of Airborne Electronic Equipment, *Tele-Tech*, **10**, 36, July, 1951; 36, August, 1951.

[2] "Proceedings of the 1956 Electronic Components Symposium," Engineering Publishers.

[3] Lucic, A., Airborne High Temperature Transformer and Reactor Components, *Aeronautical Electronics Digest*, 1955, pp. 64–79.

[4] Jacobsen, A. B., "Power Supply Characteristics and Standards for Transistorized Airborne Electronic Equipment," 1955, Publication PB-121,525 of U.S. Office of Technical Services, Washington 25, D.C.

[5] Catalog CTR-207, U.S. Office of Technical Services, Washington 25, D.C., March, 1954, lists 71 available reports on fungus proofing. Close, A. R., *et al.*, "Tropicalization of Radio and Radar Equipment: Final Report," 1949, Publication PB-114,151 of U.S. Office of Technical Services, Washington 25, D.C.

[6] MacDonald, G. J., and D. A. Thorn, Radiation from Receivers, *J. IEE (London)*, Part IIIA, No. 12, 1947, pp. 427–436.

[7] "Handbook of Preferred Circuits for Navy Aeronautical Electronic Equipment," 1955, Publication NAVAER 16-1-519, Government Printing Office, Washington 25, D.C.

Exposure to Severe Cold. Expeditions and tests in cold climates[1] have shown that under practical conditions radio, radar, and other electronic equipment are useful, but that the available models were subject to certain limitations. At one location in the Antarctic, situated on a thickness of several hundred feet of ice, local interference due to ignition and commutator noise was severe because the thickness of ice prevented the use of a ground connection. Another difficulty encountered was the unloading and moving into position of heavy diesel-engine generators for GCA systems. Other tests in severe weather showed that the handicap of extreme cold greatly reduces the efficiency with which personnel are able to install, calibrate, operate, maintain, and repair electronic apparatus.

180. Reliability of Military Equipment. The increasing amount and complexity of electronic apparatus in military and naval craft, like a chain dependent on more and more links, has often resulted in poor reliability despite compliance with the various component and system specifications. On this account special emphasis has been placed on achieving improved reliability,[2] and extensive studies and programs of action have been initiated.

While all components are important for reliability, a special need for improved quality and proper applications has been felt with regard to tubes.[3] Receiver designers should avoid such faults as dependence on tube characteristics which are uncontrolled in production, inadequate equipment operation during the latter part of normal tube life, and excessive envelope temperature. The engineering literature, as well as the applicable specifications, should be carefully followed to obtain the desired reliability (see Fig. 39).

FIG. 39. Performance vs. reliability for various classes of engineering design. (*From Leubbert, Proc. IRE, April,* 1956.)

181. Receivers for Identification and Beacon Service. Identification and beacon systems[4] include (1) an interrogator-responsor (also called an "I-R," or "challenger") which initiates the operation, and (2) a transpondor (also called a "replier" or "respondor"), which automatically answers the interrogation. An I-R is a transmitter-receiver, and a transpondor is a receiver-transmitter. Identification systems operate as accessories to radar. After the radar has shown the presence of a distant plane or vessel, the identification equipment enables the friendly character of the distant craft to be checked. All friendly craft must therefore carry transpondors. In addition, craft that are expected to make attacks on the enemy must be equipped with

[1] Bailey, H. C., Electronics in the Antarctic, *Electronics*, August, 1947, p. 82. "Military Communication Equipment Observed at Winter Tests," report based on data of M. R. Johnson, S. W. Jucius, and R. B. Zechiel, *Elec. Eng.*, August, 1947, p. 781.

[2] Henney, K., ed., "Reliability Factors for Ground Electronic Equipment," McGraw-Hill, 1956. Bromberg, B. G., and R. D. Hill, Reliability of Airborne Electronic Components, *Proc. IRE*, **41**, 513, April, 1953. Luebbert, W. F., A Systems Approach to Electronic Reliability, *Proc. IRE*, **44**, 523, April, 1956. Creveling, C. J., Increasing the Reliability of Electronic Equipment by the Use of Redundant Circuits, *Proc. IRE*, **44**, 509, April, 1956.

[3] Acheson, M. A., The Unreliable Universal Component, *Trans. IRE*, **PGRQC-9**, 49, January, 1957. Advisory Group on Electron Tubes (AGET), Tube-characteristic Changes and Equipment Reliability, *Tele-Tech and Electronic Industries*, **15**, 96, June, 1956. "Proceedings of RETMA Symposium on Reliable Applications of Electron Tubes," 1956, Engineering Publishers. Whitlock, R. S., "Techniques for Application of Electron Tubes in Military Equipment," 1955, Publication PB-111, 644-R of U.S. Office of Technical Services, Washington 25, D.C. Shipley, W. U., A Method of Measuring Cathode Interface Impedance, *IRE Convention Record*, **4**, Part 3, 64, 1956.

[4] Stuart, Lois E., Identification, Friend or Foe—Radar's Sixth Sense, *Tele-Tech*, January, 1947, p. 60. Wood, K. A., 200-Mc/s Radar Interrogator-Beacon Systems, *J. IEE (London)*, Part IIIA, No. 2, 1946, p. 481.

interrogator-responsors so as to make a check of identity before proceeding with an attack. Normally therefore interrogator-responsors are furnished in one or more of the craft in any attacking group.

The receivers of I-Rs are superheterodynes very similar to those in radar practice. The receivers of transpondors may be either superheterodynes or superregenerators with AGS.

A radar beacon system is much like an identification system. The moving craft sends out the interrogation and the beacon returns the reply; *i.e.*, the I-R is on the craft and the transpondor is at the known point where it acts as the beacon. The craft learns its location by (1) knowing the identity of the beacon from the special coding of the reply, (2) observing the location of the reply on the radar scope, and (3) taking advantage of the fact that, starting with the known location of the beacon, the reverse bearing and the same distance determine the point on the map where the craft is located.

The Brunn method of increasing the effective selectivity of a pulse receiver, described in Sec. 163, can be used advantageously in beacon service.

182. Countermeasure Receivers.[1] To reduce the value of radio and radar to the enemy, it may be necessary to (1) ascertain the frequency of the transmission and (2) radiate an interfering signal of sufficient strength and suitable modulation to prevent the enemy from realizing any benefit from his efforts. Special search receivers are required for the first of these functions. Another aspect of countermeasure-receiving technique is the use of two or more direction finders so as to ascertain the location of an enemy transmitter.

Search receivers for finding enemy radar signals[2] were developed during the Second World War, the designs being superheterodynes covering the range from 25 to 10,000 Mc. Butterfly tuners and acorn tubes were used up to about 1,000 Mc with an i-f amplifier of 3 to 4 Mc band width centered at 30 Mc. Image rejection was over 40 db at the lower frequency and over 10 db at 1,000 Mc. From 1,000 to 10,000 Mc, concentric-line oscillators with reflex klystrons were used. A 200-Mc i-f amplifier having a band width of 20 Mc at the peak and 24.5 Mc at 6 db, used with a two-cavity preselector, gave image rejections of 40 to 60 db. Reflector voltages were about 300 at the lower end of the band, increasing to 1,250 at the upper end. These equipments were air-borne and, with associated jammers, greatly reduced the effectiveness of German antiaircraft radar fire control.

Wartime countermeasure research and development centered at the Harvard Radio Research Laboratory, and the general engineering findings of this work are available in the literature.[3]

For search receivers in the centimetric band (s-h-f), it may be desired to have a large value of r-f band width so as to receive any signal in a wide range of frequencies without change of the r-f or oscillator tuning. One design approach for this purpose[4] employs simultaneously a group of heterodyne frequencies with an i-f amplifier having a substantial band width such as one-half the spacing of the oscillator components. Traveling-wave tubes may be used in such an i-f amplifier. This type of search receiver is reported to have sensitivity comparable with that of normal superheterodynes and band width comparable with that of crystal-video designs.

Panoramic equipment is of special interest in search receivers and is described below.

183. Antijamming Features in Receivers.[5] As a counter-countermeasure it is essential to design military receivers to be as immune as possible to enemy jamming.

[1] A nontechnical account of radar countermeasures is given in James Phinney Baxter, III, "Scientists against Time," Little, Brown, 1947; see pp. 158–169.

[2] Radar Countermeasures, *Electronics*, January, 1946, p. 92. Sinclair, D. B., Search Receivers for Radar Countermeasures, *Gen. Radio Experimenter*, March, 1947, p. 1. Villard, O. G., Jr., Radar Countermeasures: The Science of Immobilizing Enemy Radar, *Proc. Radio Club of America*, March, 1946, p. 7.

[3] Staff of Radio Research Laboratory, Harvard University, "Very High Frequency Techniques," McGraw-Hill, 1947.

[4] Cohn, M., and W. C. King, A Sideband-mixing Superheterodyne Receiver, *Proc. IRE*, **44**, 1595, November, 1956.

[5] Henney, K., ed., *et al.*, "Reliability Factors for Ground Electronic Equipment," McGraw-Hill, 1956; see pp. 9-12 and 9-13 on antijamming and anticlutter circuits.

The first antijamming feature is naturally against unmodulated c-w interference. Such immunity is obtainable to a good degree in a radar receiver by changing from a d-c coupling to a capacitance-resistance type between the detector and the first video stage. This prevents the first video stage being biased beyond cutoff by the rectified jamming signal.[1]

Certain types of modulation on the interfering signal can be much reduced by switching in a short-time-constant coupling (a differentiator) between the detector and the first video stage. For example, the time constant can be reduced to only 1 μsec, assuming that the desired signal consists of short pulses.

If the interference overloads the i-f amplifier at the times of high a.m. of the interference, the condition can be improved by fast a.v.c., also called "instantaneous automatic gain control" (IAGC) or "back bias."[2] Such controls can be applied to one or more of the last i-f amplifier stages.

The IAGC circuit is also useful in reducing "clutter" on the radar scope, such as due to echoes from land, sea, or clouds. Another means for reducing clutter is sensitivity-time-control (also called "STC" and "GTC"), which reduces the receiver sensitivity immediately after each transmitter pulse and then rapidly increases the sensitivity as more distant echoes arrive.

The a.f.c. of a radar receiver would be subject to jamming by c-w interference if a separate a-f-c crystal mixer were not provided.[3] By including such a separate crystal, it can be arranged to deliver the difference between the transmitter frequency and the local oscillator, so that the a.f.c. is unaffected by returning echoes or incoming interference.

Good antijamming properties and other merits have been found for a system of a-m transmission and reception in which (1) the carrier component is suppressed, (2) the two side bands are radiated, and (3) synchronous or coherent detection is employed, the reinserted carrier being developed from the two side bands.[4]

DIRECTION FINDERS

The directive properties of loops and other types of antennas make possible the design of direction finders[5] for the various portions of the radio spectrum. A direction finder is a radio receiver with a rotatable antenna or other means for determining the direction of arrival of a wave from a distant transmitter. Many instruments indicate first the direction of the infinite straight line passing through the transmitter and the direction finder, after which some further manipulation by the user is required to ascertain which end of the line is toward the transmitter, *i.e.*, to remove the 180-deg ambiguity. With an "automatic direction finder" or ADF, the needle of a dial shows the direction without ambiguity. Other types show the direction on a cathode-ray tube and are constantly sweeping the horizon so that very brief transmissions can be seen.

Direction finders with scope indicators operating in and below the v-l-f region have been found useful for locating storms over ocean areas.[6] One direction finder for this service has a tuning range from 3.6 to 17.5 kc.

[1] Ridenour, L. N., "Radar System Engineering," McGraw-Hill, 1947; see pp. 457–460.

[2] Schneider, E. G., Radar, *Proc. IRE*, April, 1946, pp. 528–578; see pp. 554–555. See also Ridenour, *loc. cit.*

[3] Straus, H. A., *et al.*, The MPG-1 Radar, *Electronics*, January, 1946, pp. 110, 114.

[4] Costas, J. P., Synchronous Communications, *Proc. IRE*, **44**, 1713, December, 1956.

[5] Keen, R., "Wireless Direction Finding," 4th ed., Iliffe & Sons, London, 1947. Bond, D. S., "Radio Direction Finders," McGraw-Hill, 1944. *J. IEE* (*London*), Part III, March, 1947, contains several articles on direction finders. *J. IEE* (*London*), Part IIIA, No. 11, 1947; two papers and a number of summaries on direction finders are given on pp. 132–168.

[6] Static D-f Reveals Storm Location, *Electronics*, February, 1946, p. 212. More on Spherics, Storm Detector, *Electronics*, April, 1946, p. 224. Adcock, F., and C. Clarke, The Location of Thunderstorms by Radio Direction-finding, *J. IEE* (*London*), Part III, March, 1947, p. 118, with discussion starting on p. 133. Kessler, W. J., and H. L. Knowles, Direction Finder for Locating Storms, *Electronics*, May, 1948, p. 106.

184. Typical Marine Direction Finders.[1] The law requires certain classes of vessels to be equipped with direction finders for the range from 285 to 515 kc. This is sufficient to operate on the marine beacons at 285 to 315 kc, the ship-to-ship compass frequency of 375 kc, and the international distress frequency of 500 kc.

The typical marine direction finder has a shielded loop antenna, a simple antenna for sense (180 deg ambiguity) determination, an r-f stage for each half of the loop (the center of the loop being grounded), a highly selective receiver, and headphones. The receiver is generally a superheterodyne of 6 or 8 tubes, and 175 kc is a typical i.f. Sensitivity values are typically 80 to 100 db below 1 volt per m (100 to 10 μv per m) for 6 mw out at 300 kc. The sharp selectivity can be seen from the representative 300-kc band widths of 6 kc at 20 db, 10 kc at 40 db, and 14 kc at 60 db.

Loop antennas for the range of 80 to 200 kc can be made smaller or more sensitive by the use of ferrite cores, and studies of rod and tube shapes,[2] and also spheroidal shapes,[3] of cores have been made.

185. Automatic Direction Finders.[4] In aircraft navigation, bearings with a direction finder are wanted on a rapid and continuously indicating basis. This need is satisfied by the ADF, which has come to be a standard type of instrument. These equipments furnish nonambiguous readings. A servo system automatically keeps the loop and an indicating needle oriented for the desired station as the plane proceeds. At least one dual type is available in which an additional antenna and receiver are provided to operate a second needle of the indicator, so that the bearings of two stations are continuously shown. The frequency coverage of ADFs is usually from 150 to 1,700 kc, and the i.f. about 112 kc. The sensitivity is about 94 db below 1 volt per m (20 μv per m). The loop antennas for ADFs are usually shielded and also housed in a specially coated radome to reduce precipitation static.

An automatic marine-type direction finder made by Raytheon for the Coast Guard[5] uses a 60-cycle switching frequency for the automatic operation. The receiver accommodates the frequency ranges 275 to 510 and 2,000 to 3,500 kc, and has an i.f. of 912 kc. The higher-frequency band includes much of the ship-shore radiotelephone service and also the international radiotelephone distress frequency of 2,182 kc.

186. Other Types. Various designs for quickly taking bearings in the h-f (3 to 30 Mc) spectrum have been made.[6] These included scope presentation of the polar type, the DF station being represented at the center and an indication in any direction showing instantly the bearing of the distant transmitter.

These equipments have two fixed antenna systems oriented at right angles to each other and connected to the two primaries of a special transformer called a goniometer. An accurate right-angle relation between the two primaries is established in the construction of the goniometer. The secondary winding rotates in the field of the two primaries and therefore picks up signal in the same way as a loop antenna rotating in space. A motor spins the secondary at high speed and also controls the scanning of the scope.

The Adcock type of antenna system, consisting only of vertical elements without the horizontal portions of the loop antenna, has desirable properties for d-f use and is

[1] Nilson, A. R., and J. L. Hornung, "Practical Radio Communication," 2d ed., McGraw-Hill, 1943; see Chap. 16, pp. 780–817.

[2] Belrose, J. S., Ferromagnetic Loop Aerials for Kilometric Waves, *Wireless Engr.*, **32,** 41, February, 1955.

[3] Wait, J. R., Receiving Properties of a Wire Loop with a Spheroidal Core, *Can. J. Technology,* **31,** 9, 1953.

[4] Scott, Allen, New Four-band ADF, *Air Transport,* April, 1946, p. 67, describes Model AN/ARN-6. "Standards and Test Procedures for Airborne Radio Equipment," RETMA Standard TR-125, July, 1951; includes direction-finding equipment. Hemphill, A., A Magnetic Radio Compass Antenna Having Zero Drag, *Proc. Natl. Electronics Conf.*, **10,** 676, 1954.

[5] Blakely, J. R., U.S. Coast Guard Automatic Direction Finder Model RD-132, *Trans. IRE,* **CS-3** (1), 16, March, 1955.

[6] Ross, W., Fundamental Problems in Radio Direction Finding at High Frequencies (3-30 Mc/s), *J. IEE (London)*, Part IIIA, No. 11, 1947, pp. 154–168. H-F Direction Finder, *Electronics,* March, 1946, p. 304. Hansel, P. C., Instant-reading Direction Finder, *Electronics,* April, 1948, p. 86.

extensively employed. Careful studies of the remaining errors in bearings taken with an Adcock system are available in the literature.[1]

An international distress frequency for lifeboats and other survival craft has been established at 8,364 kc, whence any direction finder for this general spectrum region should include this frequency.

The use of height diversity, *i.e.*, two antennas at different heights, affords improved direction-finder operation in the range of 225 to 400 Mc.[2]

Land d-f antenna systems utilizing the Doppler principle for service between 0.8 and 30 Mc offer improved accuracy.[3] The antenna equipment consists of two or three dozen individual omnidirectional antennas arranged in a large circle, each with a cable connecting to a rotating pickup switch located at the center of the circle. In effect, a single antenna is carried around the circle at the angular rate of the switch. The frequency received is normal when the antennas nearest the transmitter are effective, and likewise for those farthest away; but between these points the frequency is increased on one side and decreased on the other side because of the Doppler effect. With a switching rate of 42 cps, a circle diameter of 150 ft, and an incoming signal of 1.5 Mc, the maximum frequency deviation to one side is 30 cps.

Lobe switching, as used in radar practice, is an additional means of direction finding.[4] For this purpose the directional antenna is varied in some manner so as to make a slight change in the direction of its greatest sensitivity; at the same time the indicating means is also commutated so as to obtain an additional indication. With these facilities it is possible to direct the antenna approximately toward a "target" and then accurately adjust the direction for equality of the signal from the two lobes.

AUTOMATIC ALARM RECEIVERS

An interesting type of receiver[5] is used on vessels and at shore stations to receive automatically the international distress signal at 500 kc and sound alarm bells. The international signal for this purpose, established at the Madrid Telecommunications Conference of 1932, is a succession of dashes, each lasting 4 sec, separated by spaces of 1 sec.

187. Autoalarm Specifications. The FCC requires these receivers to operate at any frequency between 492 and 508 kc with a 30 per cent modulated signal of 100 μv. Assuming the antenna to have an effective height of 5 m, this corresponds to a signal strength of 20 μv per m. A warning must be given for various defective conditions, such as a burned-out filament. Means for checking with a local test signal must be provided. The equipment must withstand various conditions of temperature, vibration, etc.

Recent FCC regulations applicable in type tests on new autoalarm designs after the effective date of the Safety Convention (London) require that the alarm shall operate upon the receipt of "either three or four consecutive dashes when the dashes vary in length from 3.5 to as near 6 seconds as possible and the spaces vary in length between 1.5 seconds and the lowest practicable value, preferably not greater than 10 milliseconds." These provisions and also the requirement of special automatic gain characteristics are for the purpose of increasing the effectiveness of the instrument in the presence of heavy atmospherics and interference.

[1] Holbrook, J. G., An Analysis of Errors in Long Range Radio Direction Finder Systems, *Proc. IRE*, **41**, 1747, December, 1953. Travers, D. N., Spacing-error Analysis of the Eight-element Two-phase Adcock Direction Finder, *Trans. IRE*, **AP-3**, 63, April, 1955.

[2] Fantoni, J. A., and R. C. Benoit, Jr., Design of Height-diversity UHF Direction Finders, *Tele-Tech and Electronic Industries*, **15**, 90, June, 1956.

[3] Fantoni, J. A., and R. C. Benoit, Jr., Applying the Doppler Effect to Direction Finder Design, *Electronic Industries and Tele-Tech*, **16** (1), 75, January, 1957; (2), 66, February, 1957. Anderson, R. E., A Doppler Direction Finder, *Proc. Natl. Electronics Conf.*, **11**, 738, 1955.

[4] Terman, F. E., "Electronic and Radio Engineering," 4th ed., McGraw-Hill, 1955; see p. 1051 on lobe switching.

[5] Sarratt, E. R., and H. H. Abelew, Ship Autoalarm Uses Computer Techniques, *Electronics*, **30**, 180, February, 1957. Federal Communications Commission, "Rules and Regulations, Part 8, Stations on Shipboard in the Maritime Service," 1955, sold by Government Printing Office, Washington 25, D.C.

RECEIVERS FOR USE IN AIRCRAFT

188. General Requirements. The design of receivers for use in aircraft requires special attention to obtain light weight, suitable housing, resistance to vibration, adaptability to aircraft antennas, operation with available power sources, and resistance to interference conditions.[1] In addition, remote control is often required.

189. Interference Effects.[2] The ignition system of an airplane motor produces vigorous electrical effects characterized by steep wave forms which contain frequency components throughout a very wide range. These would induce an overwhelming amount of interference into any radio antenna on the plane unless they were confined by thorough shielding of the entire ignition system. In addition, shielding or filtering, or both, of most of the other wiring on the plane is required to avoid radio interference. It is especially necessary to filter leads which are connected to vibrating regulator contacts or to d-c motors. Small r-f filters are advisable in the receiver in series with each power-supply wire from outside, and also in series with external leads for audio connections, remote control, and other purposes.

Another essential element for radio reception in planes is that the various metal parts of the craft, especially moving parts, and its accessories must be firmly bonded together, as by the use of heavy copper braid.

When flying is undertaken through snow, rain, ice crystals, or dust, especially at the higher speeds, precipitation static is encountered. This is due chiefly to neutral particles striking the plane, where they are ionized by friction, furnish a high electrostatic potential to the plane, and thus cause the plane to discharge into the surrounding air and produce radio interference. The potential of the plane may rise to a quarter of a million volts or more in severe cases. The discharge which occurs is corona, sometimes known as "Saint Elmo's fire." The effect on radio reception is worst if such discharge takes place from the antenna itself, but may still be severe if the discharge occurs only from the plane itself.

An extensive investigation of precipitation static was conducted by the Naval Research Laboratory, the Aircraft Radio Laboratory, and the University of Minnesota[3] using d-c potentials as high as 1,200,000 volts. The results of this and other work have been embodied in a report of the Radio Technical Commission for Aeronautics.[4]

190. Loop Antennas for Aircraft Reception.[5] The directional properties of the loop antenna and its independence of the electric field when shielded make it a valuable

[1] "Guidance for Designers of Airborne Electronic Equipment," ARINC Report No. 403, Sept. 1, 1955, Aeronautical Radio, Inc., Washington, D.C. "Standards and Test Procedures for Airborne Radio Equipment: Procedures and Limits with Respect to Tests of Vibration, Temperature, Altitude and Humidity," EIA Standard TR-124, July, 1951. "Standards and Test Procedures for Airborne Radio Equipment; Standards for Receivers—190 Kc to 30 Mc," EIA Standard TR-125, July, 1951. Mercurio, S. P., Jr., and M. Belby, Designing Airborne Equipment for Vibration, *Electronics*, **30**, 166, January, 1957. Bromberg, B. G., and R. D. Hill, Jr., Reliability of Airborne Electronic Components, *Proc. IRE*, **41**, 513, April, 1953. Stuetzer, O. M., Transistors in Airborne Equipment, *Proc. IRE*, **40**, 1529, November, 1952.

[2] Frederick Research Corp., Bethesda, Md., "Design Techniques for Interference-free Operation of Airborne Electronic Equipment," 1952, Publication PB-111,051 of U.S. Office of Technical Services, Washington 25, D.C.; a mimeo document of 459 pp. Albin, A. L., and J. McManus, Radio Interference Control in Aircraft, *Tele-Tech and Electronic Industries*, **14**, 76, November, 1955; also in *Aeronautical Electronics Digest*, 1955, pp. 268–271. Newman, M. M., R. C. Schwantes, and J. R. Stahmann, Aircraft Radio Interference Measurements, *Elec. Eng.*, **72**, 36, January, 1953.

[3] This investigation, called the Army-Navy Precipitation Static Project, was reported in six technical papers in the *Proc. IRE*, April, 1946, p. 156P; May, 1946, p. 234. See also Dana, H. J., Corona Discharge at High Altitude and Its Control to Reduce Interference, *Proc. Natl. Electronics Conf.*, Chicago, 1947, pp. 40–46.

[4] Reduction of Precipitation Static Interference in Aircraft, Paper 68–47/DO-5 of R.T.C.A., May 19, 1947. Pelton, F. M., An Electronic System for Zeroing the Electrostatic Charge on Aircraft, *Aeronautical Electronics Digest*, 1954, pp. 183–187.

[5] Levy, G. F., Loop Antennas for Aircraft, *Proc. IRE*, February, 1943, p. 56, with correction in July, 1943, issue, p. 384. Burgess, R. E., Iron-cored Loop Receiving Aerial, *Wireless Engr.*, June, 1946, p. 172. Polydoroff, W. J., Iron-core Loop Antennas, *Electronic Inds.*, November, 1943, p. 84. Hemphill, A., A Magnetic Radio Compass Antenna Having Zero Drag, *Proc. Natl. Electronics Conf.*, **10**, 676, 1954.

type of aircraft antenna for the lower frequency ranges. Sizes up to a maximum diameter of about 12 in. have been used. Shielding is always provided, located either near the wires or as part of a streamlined housing. In case the shielding is close to the wires of the loop, the radome should be slightly conductive. Loops of 4 to 20 turns are generally used and are called "low-impedance" in distinction to designs with a greater number of turns which have been used to a limited extent. The design of the loop can be solenoidal (all turns have the same radius) or pancake (meaning that there is a gradually changing radius with all the turns in the same plane). Composite designs with both solenoidal and pancake features can be used.

Aircraft loops with iron cores have been extensively used in Europe. This design is reported to offer the advantages of smaller size and suitability for placing nearer the metal fuselage of the plane.

191. Use of 28-volt Plate Supply.[1] The 24-volt airplane battery furnishes a supply averaging about 28 volts while being charged. This is sufficient potential for the plate supply of certain tubes specially designed for the purpose. Care to avoid difficulties due to improper grid current is necessary in the operation of tubes at this low plate voltage.

The 28D7 can be operated as a low-power high-voltage supply by the provision of an oscillator circuit and a rectifier. Outputs up to 725 mw are obtainable at potentials of 50 to 250 volts, and smaller outputs are obtainable up to 500 or 600 volts. The frequency of oscillation can be located between 4 and 10 Mc. As the rectifier, the 14A4 tube can be used up to 300 volts, and the 28Z5 for higher voltages. By this use of the 28D7, limited amounts of high-voltage current can be obtained while still utilizing only a 28-volt source of power.

PANORAMIC RECEIVERS

A panoramic receiver[2] has an oscilloscope tube and suitable circuits to show the intensity of the signals available at the antenna over a frequency spectrum extending considerably above and below the frequency being received. This feature is obtained by means of an auxiliary channel which has amplifiers, a mixer, a swept oscillator, and the scope. The coupling into the auxiliary channel is from the modulator of the receiver and is made through a broad-band circuit so that signals of higher and lower frequency than the one being received will come through and appear with uniform strength on the scope.

192. Panoramic Details. The mixer of the auxiliary channel produces a new i.f. which then goes into a sharply tuned amplifier, the output of which is applied to the vertical plates of the scope. A sweep oscillator produces a saw-tooth output which is applied to the horizontal plates of the scope for scanning. The saw-tooth wave also drives a reactance tube so as to vary the frequency for which the auxiliary channel is sensitive. In this way the spectrum of interest is scanned, the frequency usually being 30 cps so as to produce a steady pattern on the scope. The horizontal axis shows the frequency at any point. Where a signal exists, an upward deflection or "pip" will be seen. If, for example, five transmissions in the frequency range of interest are taking place, five corresponding pips will appear. Furthermore, from the shape and time variation of each pip, the experienced observer can ascertain various information regarding the type of modulation and general character of the signal.

The panoramic feature is useful in communications to show frequency spaces not in use at the particular moment and therefore available for use, thus avoiding interference in crowded bands.

In military service, panoramic displays are widely used for quickly detecting hostile

[1] Tube Characteristics with 28-volt Plate Supplies, *Electronics*, December, 1947, p. 190. Hammond, C. R., E. Kohler, and W. J. Lattin, 28-volt Operation of Receiving Tubes, *Electronics*, August, 1944, p. 116.

[2] Moulic, W. E., Panoramic Principles, *Electronic Inds.*, July, 1944, p. 86. Williams, E. M., Radio-frequency Spectrum Analyzers, *Proc. IRE*, January, 1946, p. 18P. Moxon, L. A., "Recent Advances in Radio Receivers," Cambridge, 1949; see pp. 140–142 on panoramic receivers. Batten, H. W., *et al.*, The Response of a Panoramic Receiver to CW and Pulse Signals, *Proc. IRE*, **42**, 948, June, 1954.

transmissions. Also in radio operations with other friendly forces panoramic displays permit one operator to guard a much wider frequency range and respond to incoming signals more quickly.

The panoramic principle has often been applied in the form of a separate accessory or adapter to be used with a conventional receiver. One such adapter, called a "panadapter,"[1] is made in various types including one which is suitable for use with any receiver having an i.f. between 450 and 470 kc.

The broad-band coupling, which receives signals from the mixer of the receiver and serves as the input element of the panadapter, is given a double-peaked characteristic to offset the r-f selectivity of the receiver and give more uniform heights for signals of the same strength at various frequencies represented on the scope presentation.

Ferrite inductors affording a 2.25:1 inductance range under d-c control have been used in a panoramic receiver covering the frequency range from 100 to 150 Mc.[2] In this design seven such inductors have their d-c windings in series, and thus tune the oscillator and six signal-frequency couplings in the three-stage r-f amplifier. Beyond this point the circuit is a double superheterodyne with i.f.s of 30 and 3 Mc.

DIVERSITY RECEIVERS

Experience has shown that, when a signal fades at an antenna, it is usually fairly strong at a distance of 10 to 100 wavelengths away. Diversity receiving systems[3] take advantage of this fact and consist of two or more spaced antennas with special receiving equipment arranged to deliver a single output signal of improved character.

In terms of wavelength the larger separations of diversity antennas, such as 25 to 100 wavelengths, apply to the decimetric (u-h-f) and higher frequency ranges. For the dekametric (h-f) range the best values and the minimum useful values are given in Table 17.

Table 17. Values of Separation of Antennas in Dual-diversity Receiving Systems*

Frequency, Mc	Wavelength, m	Spacing of antennas, center to center, wavelengths	
		Best value	Min useful value
18.9	15.89	20	7
13.7	22	10	4
7.4	40.5	7–10	4

* Ladner, A. W., and C. R. Stoner, "Short Wave Wireless Communication," 5th ed., Wiley, 1950; see pp. 578–581.

193. Diversity Receiver Details. The conventional arrangement of receivers operating in a diversity system provides for each antenna a complete receiver up to the point where the diode detectors are connected to a common load resistor. This resistor has an audio drop across it which for radiotelephone reception is the desired signal. The common load resistor also has a d-c drop which is applied as a-v-c bias to all the receivers. With this arrangement, the d-c potential across the load resistor is determined by the strongest signal present. This is because this potential acts as a back bias on all the diodes and cuts off all of them except the one with the strongest signal. In this way only the strongest signal delivers output, and the noise that would otherwise come in from the other antennas is suppressed. As an alternative, audio mixing

[1] Popkin-Clurman, J. R., and B. Schlessel, Panoramic Reception, 1946, *QST*, March, 1946, p. 22.

[2] Gabriel, F. C., Ferrite Inductors Tune Panoramic Receiver, *Electronics*, **29**, 169, August, 1956.

[3] Jelonek, Z., E. Fitch, and J. H. H. Chalk, Diversity Reception, *Wireless Engr.*, February, 1947, p. 54. Hausman, A. H., An Analysis of Dual Diversity Receiving Systems, *Proc. IRE*, **42**, 944, June, 1954.

of signals from diversity antennas has been used followed by an audio-derived gain control.[1]

Practical points regarding receiver design for diversity operation evolved during years of experience may be found in the literature.[2]

On important circuits *frequency diversity* may be used in addition to the space diversity which has been considered so far. Frequency diversity merely consists of the transmission of the message simultaneously on more than one carrier frequency. Excellent results using 3 to 8 per cent frequency interval at about 6,000 Mc have been found by H. Magnuski of the Motorola staff and F. H. Willis of the Bell Telephone Laboratories, and Magnuski states that the possibility of using substantially lower powers may well result in spectrum conservation. However, the FCC will authorize frequency diversity only on a factual showing in each case that the required communications cannot practicably be achieved by other means (Paragraph 21.100 of their Rules and Regulations dated September, 1956).

Frequency diversity is almost synonymous with *transmitter diversity*, meaning the use of two or more transmitters to send the given message simultaneously. On shipboard, where the use of receiver diversity is normally impractical, improved service can be furnished by transmitter diversity. This has been found to give better reception of teletype signals, which are very subject to errors from fading.[3]

Somewhat related to diversity methods is the "musa" antenna system developed about 1937,[4] the term being coined from "*M*ultiple-*U*nit *S*teerable *A*ntenna." This consists of a group of rhombic antennas with phasing circuits so as to achieve (1) a high degree of vertical directivity and (2) the ability to control the vertical angle, thus "steering" the antenna. With this antenna system it is possible to favor the path on which the desired signal is arriving and discriminate against other paths. This is therefore a means of combating multipath propagation and its attendant selective fading. The musa has been found to give an improvement in quality for a considerable portion of the time and to afford a substantial improvement in signal-to-noise ratio under various conditions of poor transmission. However, the cost of the musa installation is substantial.

Diversity operation was found useful in the s-h-f range in tests at 4,500 Mc between San Diego and San Francisco.[5] At the receiving end of one of the relay links two parabolic antennas were used, one mounted 25 ft below the other. It was found that the short fades that occurred were not simultaneous at the two antennas. The outputs of the receivers were paralleled after the first video stage. The a-v-c voltages of the two receivers were connected together in the manner already described in this section. This system operated with pulse-time-modulated equipments. The 25-ft separation amounts to 112 wavelengths. The use of antennas spaced in height is called *height diversity*.[6]

The use of diversity is an important element in scatter propagation. Studies of the size for individual antennas and the separation for diversity operation[7] turn upon the distance at which the field strengths can be considered to be independent or uncorrelated. There

[1] Hamer, E. G., and D. W. Elson, An Audio-frequency Mixing System for Spaced Diversity Receivers, *J. Brit. Inst. Radio Engineers*, **13**, 123, February, 1953.

[2] Lyons, W., Criteria for Diversity Receiver Design, *RCA Rev.*, June, 1947, p. 373. Atwood, J. B., A Diversity Receiving System for Radio Frequency Carrier Shift Radiophoto Signals, *RCA Rev.*, **12**, 177, June, 1951. Sandeman, E. K., "Radio Engineering," Vol. 2, 2d ed., Chapman & Hall, 1953; see pp. 115–117. Dietsch, C. G., The Tangier Radio Relay System of RCA Communications, Inc., *Trans. IRE*, **CS-2** (1), 65, January, 1954.

[3] Hansell, G. E., Transmitter Space Diversity as Applied to Shipboard Reception, *Trans. IRE*, **CS-3** (1), 44, March, 1955.

[4] Friis, H. T., and C. B. Feldman, A Multiple-unit Steerable Antenna for Short-wave Reception, *Proc IRE*, July, 1937, p. 841; also in *Bell System Tech. J.*, July, 1937, p. 337. Polkinghorn, F. A., A Single Sideband Musa Receiving System for Commercial Operation on Transatlantic Radio Telephone Circuits, *Proc. IRE*, April, 1940, p. 157; also in *Bell System Tech. J.*, April, 1940, p. 306.

[5] Huber, G. H., Space Diversity Reception at Super-high Frequencies, *Bell Lab. Record*, September, 1947, p. 337.

[6] Egli, J. J., UHF Radio-relay System Engineering, *Proc. IRE*, **41**, 115, January, 1953.

[7] Gordon, W. E., Radio Scattering in the Troposphere, *Proc. IRE*, **43**, 23, January, 1955.

is little advantage in making a single receiving antenna larger than this because of the miscellaneous phases of the waves in the outer regions. With respect to diversity, an absence of correlation is desired, so that the separation should amount to at least this much. This distance is called the *diversity distance* or the *correlation distance.* Its value depends on whether the direction of interest is vertical, horizontal to one side, or horizontal in the direction of the ratio path. For the horizontal distance in kilometers normal to the path it is

$$D_h = 6{,}000 \frac{\lambda}{d}$$

where λ = wavelength, km
d = transmitter-receiver distance, km

The vertical diversity distance is

$$D_v = 8{,}000 \frac{\lambda}{d}$$

The horizontal diversity distance in the path direction is

$$D_a = \frac{Dv^2}{2\lambda}$$

Studies have shown that theoretically it is better not to discard the weaker diversity signals, because use can be made of them to improve the over-all signal-to-noise ratio.[1] Such a combining operation can be obtained by paralleling the outputs of all the receivers and having the output impedance of each receiver vary in proportion to the noise-power output of that receiver. In this way a receiver with a poor signal-to-noise ratio (and therefore a high noise level) will have a very high output impedance which will prevent the noise from being delivered to the output.[2]

One additional type of diversity is *polarization diversity,* meaning that the message is received on two antennas, one for horizontal polarization and the other for vertical polarization.[3]

WIDE-BAND RECEIVERS

In services such as television and pulse-type systems, receivers must possess a very much greater band width than required for simple radiotelegraph or radiotelephone operation. For example, band widths up to 4 Mc are necessary in television transmitters, and values up to 10 Mc or more characterize certain pulse transmissions.

194. Gain–Band-width Product and the Band Merit of Tubes. Studies by H. A. Wheeler[4] led to the realization that for the ideal coupling network, there is a maximum value of the product of stage gain and band width. This ideal maximum is

$$Af_w = \frac{g_m}{\pi \sqrt{C_g C_p}}$$

where A = voltage gain between equal impedances
f_w = band width, cps
g_m = transconductance of tube, mhos
C_g = total grid-to-cathode capacitance in circuit, farads
C_p = total plate-to-cathode capacitance, farads

If it is desired to realize a large part of this ideal maximum value, coupling networks of three or four reactances are necessary.

[1] Kahn, L. R., letter entitled "Ratio Squarer," *Proc. IRE*, **42,** 1704, November, 1954.

[2] Mellen, G. L., *et al.*, UHF Long-range Communication Systems, *Proc. IRE*, **43,** 1269, October, 1955. Mack, C. L., Diversity Reception in UHF Long-range Communications, *Proc. IRE*, **43,** 1281, October, 1955.

[3] Altman, F. J., and W. Sichak, A Simplified Diversity Communication System for Beyond-the-horizon Links, *Trans. IRE*, **CS-4** (1), 50, March, 1956. Glaser, J. L., and L. P. Faber, Jr., Evaluation of Polarization Diversity Performance, *Proc. IRE*, **41,** 1774, December, 1953.

[4] Wheeler, H. A., Wide-band Amplifiers for Television, *Proc. IRE*, 429, July, 1939.

For a simple parallel-resonant circuit as the coupling means, the relation takes the form[1]

$$Af_w = \frac{g_m}{2\pi C}$$

where C is the total of C_p and C_g.

This expression for the maximum gain–band-width product with a simple tuned circuit as the coupling has been used to obtain a figure of merit for comparing various tubes of interest as wide-band amplifiers. The point of view is that with wider and wider band widths, the obtainable gain with the particular tube declines until it reaches unity; *i.e.*, the output is no greater than the input. This makes A unity in the equation, so that it is only necessary to insert the values of g_m and C for the particular tube to compute its band merit. Distributed capacitances, such as those of the coil and wiring, are omitted in this calculation, as they are not properties of the tube. Table 18 gives values of the band merit for several pentode types.

Table 18. Band Merit of Various Pentode Tubes

Tube type	Transconductance, μmhos	Total of input and output capacitances, μμf	Band merit, Mc
6K7	1,600	20	13
1AD4*	1,700	8	36
5902*	4,200	14	48
6AU6WA	5,000	12	70
5749/6BA6W	4,300	11	70
5639*	9,000	17	85
6AG7	11,000	20	88
6AC7	9,000	16	90
5899*	4,500	7.7	94
6AG5	5,100	8.3	98
6AH6	8,000	13	98
5840*	5,000	7.6	105
5654/6AK5W	5,100	6.8	120
5847/404-A	12,500	9.5	210

* Subminiature type.

An essential part of a wide-band receiver is the i-f amplifier. A general treatment of this is given in Sec. 77. Stagger tuning and feedback methods are used to obtain wide-band response.

After detection of the signal, it is usually necessary to amplify it further. In wide-band receivers such an amplifier is usually called a "video amplifier" because the signal is in many cases displayed on a scope as a television picture or as a radar pattern. For application to television see Chap. 22 of this volume and references;[2] for radar applications see references[3] and Secs. 195 and 196.

[1] Ford, G. T., Characteristics of Vacuum Tubes for Radar Intermediate Frequency Amplifiers, *Bell System Tech. J.*, July, 1946, pp. 385, 388–390.

[2] Fink, D. G., "Principles of Television Engineering," McGraw-Hill, 1940; see pp. 207–260. Video Amplifier H.F. Response, *Wireless World*, September, 1946, p. 301; October, 1946, p. 333. Schlesinger, Kurt, Low-frequency Compensation for Amplifiers, *Electronics*, February, 1948, p. 103. Larsen, M. J., Low-frequency Compensation of Video-frequency Amplifiers, *Proc. IRE*, October, 1945, p. 666. Wheeler, H. A., The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *Proc. IRE*, June, 1939, p. 359. Kallmann, H. E., R. E. Spencer, and C. P. Singer, Transient Response, *Proc. IRE*, March, 1945, p. 169; includes many computed curves of response to unit-step signal. Sziklai, G. C., and A. C. Schroeder, Cathode-coupled Wide-band Amplifiers, *Proc. IRE*, October, 1945, p. 701.

[3] Schneider, E. G., Radar, *Proc. IRE*, August, 1946, p. 528; see pp. 551–556 for treatment of receivers. Berkner, L. V., Naval Airborne Radar, *Proc. IRE*, September, 1946, p. 671. Ridenour, L. N., "Radar System Engineering," McGraw-Hill, 1947. Fink, D. G., "Radar Engineering," McGraw-Hill, 1947. Staff of Radar School MIT, "Principles of Radar," 2d ed., McGraw-Hill, 1946. Morrison, L. W., Jr., The Radar Receiver, *Bell System Tech. J.*, October, 1947, p. 693. Miller, S. E., Considerations in the Design of Centimeter-wave Radar Receivers, *Proc. IRE*, April, 1947, p. 340. *J. IEE* (*London*), Part

The discussion of the product of gain and band width given above applies to the case of one tube in each stage. In distinction, two or more tubes per stage can be used by either (1) the method of *distributed amplification*,[1] or (2) the use of tubes in parallel with filters to divide the frequency band among the tubes of the stage. Either method permits wider bands to be amplified than is possible with a single tube in each stage. Fundamental studies by Wheeler[2] have resulted in the conclusion that for any particular type of tube there is a maximum value of the *speed of amplification*, which is defined as the quotient of the gain divided by the delay time incurred in amplifying the signal by this amount of gain. For the 6AK5 pentode this maximum value is 0.65 db per millimicrosecond.

Promising results have been obtained in studies and tests of transistor wide-band amplifiers.[3]

RADAR

195. General Properties. Specifications of a number of radars used in the Second World War are available in the literature.[4] Radars made for installation on the ground included models with carrier frequencies from 195 to 10,000 Mc, pulse durations from ¼ to 30 μsec, pulse repetition rates from 200 to 4,098 pulses per sec, receiver noise figures from 7 to 18.5 db, and band widths from 1.25 to 10 Mc.

Air-borne radars included models operating at carrier frequencies from 2,500 to 10,000 Mc with pulse durations from 0.4 to 1.13 μsec, pulse repetition rates from 400 to 2,000 pulses per sec, receiver noise figures from 11 to 21 db, and receiver band widths from 1.9 to 8 Mc.

The postwar importance of radar led the FCC to assign the following three frequency bands for general maritime radar use: 3,000 to 3,246, 5,460 to 5,650, and 9,320 to 9,500 Mc. These bands accommodate the war designs in the neighborhood of 3,000 and 9,000 Mc and, in addition, establish the new 5,460 to 5,650 band at which experience can be gained. Radars for higher frequencies have better antenna directivity but suffer greater attenuation due to rain. The choice of the best operating frequency is therefore a compromise.

The receiver of a radar is essentially a wide-band type, as discussed in Sec. 194, to which a cathode-ray display has been added. Various auxiliary circuits for the display, and often others for the receiver proper, are required to give the necessary performance.[5]

Various types of displays are available. Maritime equipments normally use the PPI (plan position indicator), a maplike presentation with the observer's location at the center. Military designs use the PPI and various other types.[6]

As typical receiver details in maritime radars,[7] the following can be mentioned:

IIIA, No. 1, 1946, pp. 1115–1158; this issue gives four papers on radar receivers. Van Voorhis, S. N., "Microwave Receivers," McGraw-Hill, 1948. Other McGraw-Hill volumes related to receivers are Hamilton, Knipp, and Kuper, "Klystrons and Microwave Triodes"; Torrey and Whitmer, "Crystal Rectifiers"; Pound, "Microwave Mixers"; Chance *et al.*, "Electronic Time Measurements"; Soller, Starr, and Valley, "Cathode Ray Tube Displays"; Lawson and Uhlenbeck, "Threshold Signals."

[1] Stockman, Harry, "Introduction to Distributed Amplification," 2d ed., published by SER Co., Waltham, Mass., 1956. Ginzton, E. L., *et al.*, Distributed Amplification, *Proc. IRE*, **36**, 956, August, 1948.

[2] Wheeler, H. A., "The Maximum Speed of Amplification in a Wideband Amplifier," No. 11 in volume "Wheeler Monographs," published by Wheeler Laboratories, Great Neck, N.Y., 1953.

[3] Brunn, G., Common-emitter Transistor Video Amplifiers, *Proc. IRE*, **44**, 1561, November, 1956. Angell, J. B., Wide Band Amplification with Surface-barrier Transistors, *IRE Convention Record*, **2**, Part 7, 15, 1954.

[4] Radar Specifications, *Electronics*, November, 1945, p. 116. Airborne Radar Specifications, *Electronics*, February, 1947, p. 132.

[5] Morrison, L. W., Jr., The Radar Receiver, *Bell System Tech. J.*, **26**, 693, October, 1947; also published as pp. 364–487 of book "Radar Systems and Components" by staff of Bell Telephone Laboratories, Van Nostrand, 1949. Bowen, E. G., ed., "A Textbook of Radar," 2d ed., Cambridge, 1954; see pp. 321–508 on receiving topics.

[6] Westman, H. P., ed., "Reference Data for Radio Engineers," 4th ed., International Telephone and Telegraph Corp., 1956; see pp. 805–807.

[7] Moore, C. E., Low Cost "Big Radar" for Small Vessels, *Trans. IRE*, **CS-3** (1), 14, March, 1955. Isbister, E. J., New Marine Radar, *Trans. IRE*, **CS-3** (1), 31, March, 1955.

operating frequency, approximately 9 kMc; STC (sensitivity time control) provided; over-all band width adjustable to 4 or 12 Mc, or fixed at 8 Mc; and PPI display of 10 to 16 in. diameter.

196. Receiver Design Matters. One of the methods of stating the sensitivity of radar receivers is by their *tangential sensitivity*. This is measured, as shown in Fig. 40, by determining the amplitude of pulse such that its base line is approximately tangent to the upper edge of the noise signal. Various observers have found that this corresponds to a signal-to-noise ratio of about 8 db.[1]

The over-all band width of a radar receiver can be made about equal to the reciprocal of the pulse duration for good detectability of weak signals. For accurate range determination the band width can advantageously be increased to as much as four times this value.

For holding and tracking a particular target, the radar receiver can be provided with suitable gating (to isolate the corresponding echo signal) and with a.v.c. to maintain the desired signal amplitude. In such cases satisfactory a-v-c action without low-frequency noise otherwise encountered has been obtained by providing an a-v-c filter circuit having a quick attack and a slow release time.[2]

FIG. 40. Appearance of oscilloscope in use of tangential pulse signal for measurement of the selectivity of a pulse receiver.

Radar receivers are often provided with STC (sensitivity time control), which operates automatically to reduce the sensitivity immediately after each transmitted pulse and then increase the sensitivity steadily as the more distant, and therefore weaker, echoes are received. An alternative design is a logarithmic receiver, whose output is proportional to the logarithm of the input.[3] Either STC or a logarithmic feature operates as an anticlutter means by removing the saturation from fixed objects or sea return which otherwise may obscure desired echoes at close and medium distances.

Other anticlutter methods have much in common with antijamming circuits (Sec. 183). One method is fast-acting a.v.c., often called IAGC for *instantaneous automatic gain control*, which lowers the gain promptly on the receipt of strong echoes or interference. Another method is FTC, or *fast-time-constant*, circuit, which differentiates the wave form and is useful for reducing clutter due to rainfall. Circular polarization is also beneficial in combating rainfall clutter.

Two other methods of improving radar reception have been used, especially in military service. One of these is MTI, or the *moving-target indicator*, which operates by a dual-channel feature and a delay line in the receiver so that fixed echoes are canceled. The other method of improving radar reception involves the distant object rather than the radar receiver and consists of providing at the distant point or craft a *transpondor*, which is an automatic receiver-transmitter. This offers the advantages of much stronger signals and also the means of furnishing identification information regarding the distant object.

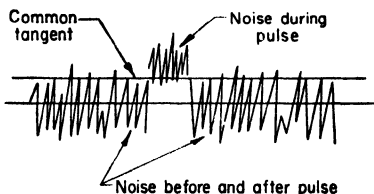
Miniaturized i-f amplifiers suitable for radar service are described at the end of Sec. 130.

The proportioning of the second-detector circuit of the radar receiver can be facilitated by a nomogram prepared by C. W. Young.[4]

[1] Van Voorhis, S. N., ed., "Microwave Receivers," Vol. 23, M.I.T. Radiation Laboratory Series, McGraw-Hill, 1948; see pp. 292, 456.

[2] Delano, R. H., and I. Pfeffer, The Effect of AGC on Radar Tracking Noise, *Proc. IRE*, **44**, 801, June, 1956.

[3] Croney, J., A Simple Logarithmic Receiver, *Proc. IRE*, **39**, 807, July, 1951. Raven, F. A., New Type Instantaneous Logarithmic Wide-band Amplifier, *Technische Mitteilungen PTT*, No. 5, 1951; available in English translation as Publication PB-121,485 from U.S. Office of Technical Services, Washington 25, D.C.

[4] Young, C. W., Radar Second Detector Filter Nomograph, *Electronics*, **29**, 188, May, 1956.

SPECIAL RECEIVERS

Numerous cases arise in which designs of radio receivers for special purposes are needed. A few such instances illustrating the wide range of requirements are given.

197. Field-strength Meters.[1] Various special features are necessary in receivers used for the measurement of field strength. Precise antenna dimensions, suitable calibrating arrangements, very stable general operation, an indicating instrument, and switching to choose the detector time constant are some of the requirements. Field-strength meters are used for the measurement of both signal strength and interfering disturbances.[2]

Field-strength meters for the v-h-f (metric) and part of the u-h-f (decimetric) bands were covered by a proposed American standard[3] issued in 1952. An instrument for the 600-Mc region was designed at the National Physical Laboratory in England, using a silicon crystal and an i-f amplifier operating at 30 Mc.[4]

198. Receivers for Radio Astronomy.[5] An outstanding aspect of the reception and study of radio waves from heavenly bodies and regions is the very low signal strength of the radiation of interest in comparison with the general noise background. As a result, the signal-to-noise ratio is far less than unity. Special methods are therefore necessary, usually consisting of rapid switching (*e.g.*, at a 500-cps rate) from the signal of interest to some suitable reference condition so that the result of many such comparisons is obtained.

Frequencies of the entire normal radio spectrum are received as cosmic radiation, as discussed in Sec. 25. One frequency is 1,420 Mc, associated with hydrogen, and 327 Mc, associated with deuterium (heavy hydrogen).

A compact astronomical radio receiver can act as a sextant for nautical use in shooting the sun in cloudy weather.[6]

199. CONELRAD Receivers. The FCC has established the CONELRAD (control of electromagnetic radiation) system as a civil defense aid in the event of an enemy airplane attack. One feature is that certain key broadcast stations receive the official notice, make an announcement, and go off the air. Many other broadcasting and other classes of stations are required to monitor one of these key stations, and upon an alert likewise to leave the air. A CONELRAD receiver[7] is one designed to perform this monitoring, and includes the following features: (1) An r-f channel tuned to the key station but not usually connected to the loud-speaker; (2) a relay whose armature is held up by the detected d.c. of the key-station signal; and (3) an alarm actuated by the release of the relay armature or by failure of the set on the fail-safe principle. In addition, the set may include another r-f channel, or a change-of-tuning feature, so as to reproduce on the loud-speaker the civil-defense instructions received on 640 or 1,240 kc, the channels assigned for this purpose.

[1] Wind, M., ed., "Handbook of Electronic Measurements," Polytechnic Institute of Brooklyn, 1956; see Chap. 8.

[2] U.S. Navy Specification MIL-I-16910-A, "Interference, Measurement, Radio, Methods and Limits: 14 Kilocycles to 1000 Megacycles," Aug. 30, 1954.

[3] "Proposed American Standard Specifications for Radio Noise and Field Intensity Meters, 20 to 1000 Megacycles/Second," Publication C 63.3 of American Standards Association, 1952; also designated NEMA 131-1952 and EIA Engineering Bulletin 41.

[4] Gordon-Smith, A. C., 600-Mc/s Field-strength Meter, *Wireless Engr.*, **29**, 306, November, 1952.

[5] Lovell, B., and J. A. Clegg, "Radio Astronomy," Wiley, 1952. Southworth, G. C., Early History of Radio Astronomy, *Sci. Monthly*, **82**, 55, February, 1956. Four papers on radio astronomy, *Proc. Natl. Electronics Conference*, **11**, 496, 1955. Kraus, J. D., and E. Ksiazek, New Techniques in Radio Astronomy, *Electronics*, **26**, 148, September, 1953. Aarons, J., Antenna and Receiver Measurements by Solar and Cosmic Noise, *Proc. IRE*, **42**, 810, May, 1954. Pederzani, T., German Radio Observatory Stockert, *Elec. Eng.*, **76** (3), 196, March, 1957. Muller, C. A., A Receiver for the Radio Waves from Interstellar Hydrogen, *Philips Tech. Rev.*, **17**, 305, May, 1956; 351, June, 1956.

[6] McCoy, D. O., An All-weather Radio Sextant, *Navigation*, **4**, 309, December, 1955; also in *IRE Convention Record*, **3**, Part 5, 92, 1955.

[7] Carey, R. B., Automatic Alarm Device for CONELRAD Radio Alerting, *Elec. Eng.*, **73**, 963, November, 1954. Quenstedt, R. E., Conelrad Receiver with Built-in Alarm, *Electronics*, **27**, 156, January, 1954. Mitchum, M. M., CONELRAD Alert Receiver, *Electronics*, **25**, 110, October, 1952.

200. Receivers with Privacy Feature.[1] For public radiotelephone service it is usual to provide some form of speech scrambling so that eavesdropping with simple types of receivers is prevented. One type of privacy equipments, described in Sec. 177, utilizes division of the voice-frequency range into several portions which are then exchanged, with or without inversion of the individual divisions. A simpler method of scrambling is to invert the entire speech band as a single unit, the usual practice being to have 1,550 cps as the center point. In this case, 1,000 cps, for example, becomes 2,100, and 2,100 becomes 1,000; similarly, 500 cps and 2,600 change places; likewise, 100 and 3,000 cps.

Warbling of the carrier frequency is also available as a privacy method. The extent of the warbling must obviously be kept within bounds in order to avoid excessive increase of the frequency spectrum.

201. Receivers Using Synchronized Added Carrier. Experimental designs have been made of receivers using added synchronized carrier at the signal detector.[2] The "Synchrodyne" receivers are of this type and include circuits for automatically maintaining synchronism of the local oscillator. For special conditions, such as when the noise exceeds the signal, the use of added synchronous carrier has been found advantageous.

Promising characteristics have been found with a system of suppressed-carrier double-side-band transmission in which the receiver recreates the carrier using only the information in the two side bands.[3] This system therefore requires no pilot carrier for its operation.

UNDERWRITERS' REQUIREMENTS

Numerous provisions for the reduction of shock and fire hazards have been established by Underwriters' Laboratories, which is sponsored by the National Board of Fire Underwriters. Receiver models examined and found to comply with these provisions are privileged to bear a label showing approval by Underwriters' Laboratories. The latest edition of the "Standard for Radio and Television Receiving Appliances" of Underwriters' Laboratories will be furnished to receiver design engineers on request to any one of their offices in New York, Chicago, or Santa Clara, Calif. This standard is a pamphlet giving the specific electrical and mechanical requirements for the approval of equipment, and covers radio receivers, television receivers, phonographs for home use, and recorders.

202. Receiver Hazards. With regard to fire hazard,[4] one of the most important requirements for a-c receivers is that the power transformer must be completely housed in noncombustible material without louvers or holes except as required for the entry of wires. The purpose of the housing is to prevent the escape of flame or molten metal in case the transformer is overloaded by the short circuit of some secondary winding.

With a-c–d-c sets, the chief fire hazard is the first electrolytic filter capacitor, which for a-c operation and a short circuit of the rectifier tube is subjected to a severe condition. The connection of the filter circuit to a tap on the rectifier-tube filament is advantageous in reducing this hazard.

Shock hazard includes exposed conducting parts such as antenna lead, ground lead, metal cabinet if used, phono tone arm if used, and shafts and setscrews of control knobs.

[1] Reich, H. J., "Theory and Applications of Electron Tubes," 2d ed., McGraw-Hill, 1944; see pp. 299–300. Shannon, C. E., Communication Theory of Secrecy Systems, *Bell System Tech. J.*, **28**, 656, October, 1949.

[2] Smith, R. A., The Relative Advantages of Coherent and Incoherent Detectors: A Study of Their Output Noise Spectra under Various Conditions, *Proc. IEE (London)*, Part III, **89**, 401, September, 1951. Tucker, D. G., The Synchrodyne and Coherent Detectors, *Wireless Engr.*, **29**, 184, July, 1952; see also summary No. 2883 on p. A214. Battersby, L. R., and E. A. Conover, "Investigation of the Synchrodyne," Report M-1478 of Signal Corps Engineering Laboratories, Ft. Monmouth, N.J., 1952; available from U.S. Office of Technical Services, Washington 25, D.C., as Publication PB-116,021.

[3] Costas, J. P., Synchronous Communications, *Proc. IRE*, **44**, 1713, December, 1956.

[4] Heaton, H. T., Preventing Fires from Electrical Causes in the Design and Manufacture of Radio and Television Receivers, *Trans. IRE*, **BTR-1** (2), 28, April, 1955.

Various requirements must be satisfied by flexible cords with regard to (1) character of wire, (2) bushing or equivalent treatment of the hole, (3) strain relief to prevent a pull on the cord being transmitted to internal connections, and (4) means to prevent the cord from being pushed into the hole.

Temperature limits for various materials inside the receiver must not be exceeded. Each wire must be suitable for the voltage, temperature, and other conditions prevailing.

A-c–d-c receivers with the chassis directly connected to one side of the power mains are permitted under certain conditions including the provision of a substantial back cover and interlock. In the case of f-m receivers for a-c–d-c operation, this has been found the most practical arrangement because it avoids the regeneration difficulties encountered with an isolated chassis.

AUTOMATIC MANUFACTURE OF RECEIVERS

203. General Types. The designer of receiver circuits, and especially of the associated mechanical features, must consider the manufacturing methods to be used. These methods include techniques adapted for automatic production,[1] especially the various forms of "printed wiring" and the modular type of construction.

The term *printed wiring*[2] is applied to any construction which produces flat wiring on a supporting board of insulation such as paper-laminated phenolic. The various methods include photographic processes, chemical etching, embossing, and other operations. The use of printed-circuit construction saves the labor otherwise needed to make the many individual connections between parts. It also affords more compact assemblies, which is often an important advantage. Further treatment of miniaturization is given in Sec. 130.

The modular type of construction for automatic manufacture[3] utilizes assemblies of ceramic wafers or disks. Each module consists of several of these disks slightly separated one above the other in a stack and provided with both mechanical support and electrical connections by means of straight bare wires going from disk to disk. Notches are provided in the edges of the disks for the wires. Each disk provides or supports one or more circuit elements, such as a capacitor or resistor, and the top disk can support a tube socket.

COLOR CODES FOR WIRING

204. EIA Code. Standard REC-108-A of EIA[4] establishes a color code for chassis wiring which includes basic functions according to body wire color and subordinate

Table 19. Body or Solid Wire Colors According to Basic Function*

Color	Function
Black	Grounds, grounded elements, and returns
Brown	Heaters or filaments, off ground
Red	Power-supply B plus
Yellow	Cathodes
Green	Control grids
Blue	Plates
White	Above- or below-ground returns, a.v.c., etc.

* From Standard REC-108-A.

classification according to tracer color. The body colors and their basic significance are given in Table 19. Other color codes will be found in other chapters of this volume.

[1] Carlzen, C. F., and W. L. Lehner, Automatic Assembly of a Radio Set, *Sylvania Technologist*, **9**, 46, April, 1956. Proc. of Symposium on Printed Circuits, Engineering Publishers, 1955. Henney, K., ed., "Reliability Factors for Ground Electronic Equipment," McGraw-Hill, 1956; see Chap. 10. Special issue of *Tele-Tech and Electronic Industries*, **14** (12), December, 1955.

[2] Markus, J., Mechanized Production of TV Wiring Boards, *Electronics*, **30**, 138, Apr. 1, 1957. Place, S. W., An Up-to-date Appraisal of Laminated Plastics for Printed Circuits, *Insulation*, April, 1956. Gerhold, R. A., Printed Circuits and the Automatic Factory, *Proc. Natl. Electronics Conference*, **8**, 481, 1952.

[3] Henry, R. L., C. C. Rayburn, and J. G. Black, Mechanized Production of Electronics, Publication PB-111,278 of U.S. Office of Technical Services, Washington 25, D.C., 1954; this is an 85-page mimeo report on the "Tinkertoy" modular development. Henry, R. L., and C. C. Rayburn, Mechanized Production of Electronic Equipment, *Electronics*, **26**, 160, December, 1953.

[4] "Color Coding," Standard REC-108-A, 1949.

It is also specified that, when antenna and ground connections are provided on the receiver, the antenna lead shall be blue and the ground lead black. In the case of leads for special antenna connections, combinations of blue and black are specified.

ANTENNAS FOR BROADCAST RECEPTION

A general treatment of antennas is given in Chap. 20, and loop antennas are discussed in Sec. 31 *et seq.* of the present chapter. The present section gives data on some open-type antennas and systems especially designed for broadcast reception.

205. F-m Antennas. Use of the home electric wiring as an antenna for f-m signals on the 100-Mc band has been found practical. This form of antenna, although not so good as a well-installed outside dipole, has been found an acceptable substitute and permits f-m table-model receivers to operate without the necessity of installing an outside antenna. A minor difficulty is that the results are sometimes poor at one outlet while they are good at another outlet in the same room.

In the 100-Mc f-m band extensive use has been made of folded-dipole antennas constructed of 300-ohm balanced transmission line and mounted in the top of the cabinet. Such an antenna extending across the width of a 40-in. cabinet and extending along the sides for 9 in. had an efficiency within 1 db of a good dipole used as reference standard, and had voltage standing-wave ratios of about 5 at 88 Mc, 1.5 at 98 Mc, and 3.5 at 108 Mc. A more compact design, extending across the 24-in. width of a cabinet and continuing for the full length of the 12-in. sides and then for 8 in. around the corner, thus having a total length of 64 in., had an efficiency of −2.5 db with respect to the same reference standard, and voltage standing-wave ratios of 6, 3.6, and 7, respectively, at the three frequencies already mentioned.

206. Master Antennas. Technique is available for the design of master antenna systems suitable for supplying many outlets, such as in apartment houses, and providing service on the broadcast band, the h-f international and amateur bands, and the lower television channels.[1]

Installation of antennas for furnishing 100-Mc f-m service, as well as h-f and m-f reception, have been made on a large scale in some apartment-house developments.[2] One antenna in these installations supplies one antenna coupler which delivers signal to one or two transmission lines consisting of RG-59/U flexible coaxial cable. Each of these cables supplies as many as 18 individual outlets. In this installation the outlets are directly in parallel across the transmission line; *i.e.*, no receiver couplers are used. Each run of RG-59/U is terminated at the end with a 100-ohm resistor. A polarized outlet is used in each apartment to prevent a ground being placed on the center conductor of the RG-59/U. A power outlet for the 117-volt supply of the radio receiver is provided conveniently at the same location.

General References

207. Books on Design. The following volumes give general treatments of receiver design:

Dawley, R. L., editor: "The Radio Handbook," 14th ed., Editors and Engineers, Summerland, Calif., 1956.

Electronics Training Staff of the Cruft Laboratory, Harvard University: "Electronic Circuits and Tubes," 948 pp., McGraw-Hill, 1947. Radio receivers are discussed in Chap. 23, pp. 741–782.

Langford-Smith, F.: "Radiotron Designer's Handbook," 4th ed., RCA, Harrison, N.J., 1952.

Moxon, L. A.: "Recent Advances in Radio Receivers," Cambridge, 1949.

Philips Technical Library, Series on Electronic Valves, Books IIIB, IV, V, and VI, English translations available from Elsevier Press, Houston 25, Tex.

"Radio Amateur's Handbook," 1958 ed., American Radio Relay League, West Hartford, Conn.

"RCA Receiving Tube Manual RC-17," 320 pp., Tube Dept., RCA, Harrison, N.J. Contains considerable material on general characteristics and uses of tubes.

---

[1] Aceves, J. G., Recent Improvements in Master Antenna Systems, *Proc. Radio Club of America*, July, 1940, p. 1. Urrico, F. X., Jr., A Standard Method for Determining Multicoupler Sensitivity Loss, Publication PB-111,541 of U.S. Office of Technical Services, Washington 25, D.C., 1954. Long, W. G., Jr., Linear Coupling Amplifier, Publication PB-115,138 of U.S. Office of Technical Services, Washington 25, D.C., 1954, reports a multicoupler study.

[2] Klingenschmitt, F. A., FM Antennas for 12,497 Families in Metropolitan Life Projects, *FM & Television*, May, 1947, p. 21.

Strutt, M. J. O.: "Modern Ultra- and Extreme-shortwave Reception," 387 pp., Van Nostrand, New York, and Macmillan, London, 1947.

Sturley, K. R.: "Radio Receiver Design," Part I, 2d ed., 667 pp., Wiley, 1953. Covers r-f amplification and detection. Part II, 480 pp., published by Chapman and Hall, London, 1945, covers audio amplification, power supply, automatic gain control, special tuning methods, receiver measurements, f-m reception, and television reception.

Van Voorhis, S. N.: "Microwave Receivers," 618 pp., McGraw-Hill, 1948.

Zepler, E. E.: "The Technique of Radio Design," 2d ed., Wiley, 1951.

208. Terminology in British References. The following glossary of British radio terms may be useful when consulting the various excellent books and journals published in Great Britain:

Accumulator = storage battery
Bush = bushing
Decoupling = use of filter to prevent feedback coupling
Earth = ground
E.H.T. = extra high tension = plate supply for cathode-ray tube
Former = coil form
Four-pole = four-terminal passive network
Gramophone = phonograph
H.T. = high tension = plate supply
Long-wave band = 900–2,000 m = 333–150 kc
L.T. = low tension = filament supply
Mains = power wires in the house for lighting and appliances
Medium-wave band = 200–550 m = 1,500–550 kc
Picofarad = pf = micromicrofarad
Radiogram = phonoradio combination
Relay = furnishing of programs over wires to homes or apartments from a master radio receiver or by wire from the studio of a radio station
Screening = shielding
Shroud = strap for holding transformer laminations
Speech coil = voice coil (of a loud-speaker)
Tag = lug
Torch = flashlight
Valve = tube
Valve holder = tube socket

CHAPTER 20

ANTENNAS

By John N. Hines[1]

The transmission and reception of electromagnetic waves used for radio communication are accomplished by radiators and collectors exposed in space and known as antennas. In this respect, an antenna acts as an efficient transformer between free space and a transmission line, or vice versa. Between the transmitting and the receiving antennas there is a combination of earth, water, air, and ionospheres which constitute the mediums in which electromagnetic waves are propagated. The action of the waves in traversing these mediums is very complex, being dependent upon many known and other unknown factors. Prominent among the known factors are the transmitting frequency, the radiation characteristics of the transmitting antenna, the orientation of the path of transmission in the earth's magnetic field, the time of day and the conditions of daylight and darkness along the path, the season of the year, the solar activity, the electrical characteristics of soil or water in the immediate vicinity of the antenna as well as along the path of the surface waves, the immediate conditions of ionization of the atmosphere at various levels, the distance between transmitter and receiver, and the characteristics of the receiving antenna.

The whole field of antenna technology has grown from the striving to launch and collect radiant energy in the most efficient ways dependent upon the empirical facts of wave propagation. If the earth were a smooth perfectly conducting sphere in a vacuum, antenna techniques, except for size, would not vary much over the entire useful range of radio frequencies. In reality, the lowest frequencies are best utilized by ground-wave transmission. These gradually blend into the frequencies where both ground waves and sky waves have separate usefulness. Gradually these blend into the frequencies for which ground waves are rendered useless and only sky waves can be usefully employed. Eventually frequencies are reached where the ionospheres no longer reflect waves and where range is limited to the diffraction region beyond the horizon. At still higher frequencies, this diffraction zone is reduced to virtual uselessness, and transmission is confined to purely optical circuits, but complicated sporadically by meteorological effects which produce mirage. Meteorology plays a greater part in wave propagation as the frequencies enter the s-h-f region. In the frequency range from beyond the point where radio waves no longer reflect from the ionosphere to the s-h-f region exists a recently discovered propagation phenomenon called *forward scatter* which greatly extends the limits of long-range communications. The different portions of the spectrum each require different basic antenna techniques, and in each portion a large variety of antenna structures has been evolved with different degrees of general or special utility.

It must be emphasized that antenna applications are dependent upon the nature of wave propagation for any frequency and transmission circuit, and for that reason the antenna engineer must also be familiar with propagation physics.

It is impossible to place on paper all the various considerations that determine the design details of an antenna. For different applications, different weights must be placed on each factor in the complex combinations of electrical, mechanical, and economical aspects of the design. Even though the same fundamental principles apply

[1] Antenna Laboratory, Ohio State University. On low- and medium-frequency antennas, the author has drawn to a considerable extent on the material on these subjects appearing in the fourth edition and prepared by E. A. Laport.

generally, a carefully engineered antenna system is an empirical entity, calling for a certain seasoned judgment gained only from experience or close study of the theory and techniques developed for various purposes throughout the years. But there is no intrinsic mystery to the subject, and the fine points merely provide those extra few decibels of superiority; and whether this superiority is worth what it sometimes costs, the designer must decide for the particular case. As in most fields, overdesign may demonstrate poor judgment as much as underdesign.

In general, it is possible to excite any antenna at any frequency simply by applying enough potential to drive a current against the impedance it presents at that frequency, whatever it may be, or to match this impedance into that of the generator so that all the power is delivered up to the point of termination. Beyond this point, the power must either be consumed in internal losses in the system or be radiated. But to radiate the power in a particular manner into space for effective utilization, the antenna must be designed to have current or aperture distributions in the radiator which shape the radiated field in a desired, as distinguished from a random, manner.

1. Definition of Terms[1]

Antenna (Aerial). A means for radiating or receiving radio waves.

Unipole or Isotropic Antenna. A hypothetical antenna radiating or receiving equally well in all directions. In electromagnetics, unipoles do not exist physically but represent convenient reference antennas for expressing the directive properties of actual antennas.

Electric Dipole (Doublet). A pair of equal and opposite charges an infinitesimal distance apart. In electromagnetics, the term "dipole" is often applied to two equal and opposite oscillating charges an infinitesimal distance apart; in this sense, it is synonymous with an electric-current element or doublet.

Antenna Array. A system of antennas coupled together for the purpose of obtaining directional effects.

Omnidirectional Antenna. An antenna producing essentially constant field strength in azimuth and a directive radiation pattern in elevation.

Unidirectional Antenna. An antenna which has a single well-defined direction of maximum radiation.

Radiating Element. A basic subdivision of an antenna array which in itself is capable of radiating or receiving r-f energy.

Parasitic Element. A radiating element, not coupled directly to the feed line of the antenna array, which materially affects the pattern of the antenna.

Dipole Antenna. A straight radiator, usually fed in the center and producing a maximum of radiation in the plane normal to the axis. A half-wave ($\lambda/2$) linear radiator is usually called a half-wave dipole. A vertical dipole antenna will radiate only theta (θ) polarization.

Loop Antenna. An antenna consisting of one or more complete turns of conductor and functioning by virtue of the circulatory current therein. A loop antenna with uniformly distributed in-phase current is sometimes referred to as a *magnetic dipole.* Such a horizontal loop antenna will radiate only phi (ϕ) polarization.

Aperture (of a Unidirectional Antenna). That portion of a plane surface near the antenna, perpendicular to the direction of maximum radiation, through which the major part of the radiation passes.

Aperture Illumination. The field distribution in amplitude and phase over the aperture.

Antenna Resistance. The power supplied to the entire antenna circuit divided by the square of the effective (rms) antenna current referred to a specified point. Antenna resistance is made up of such components as radiation resistance, ground resistance, r-f resistance of conductors in the antenna circuit, and the equivalent resistance due to corona, eddy currents, insulator leakage, and dielectric power loss.

[1] Most of these definitions are taken from (1) IRE Standards on Antennas, Modulating Systems, and Transmitters: Definitions of Terms, 1948, (2) IRE Standards on Antennas: Methods of Testing, 1948, and (3) IRE Standards on Antennas and Waveguides: Definitions of Terms, 1953.

Radiation Resistance. The power radiated by the antenna divided by the square of the effective (rms) antenna current referred to a specific point, usually where power is introduced.

Self-impedance. The ratio of the impressed voltage and the current at the feed point of a single radiating element in the absence of any influences from other radiators or ground.

Mutual Impedance. The mutual impedance Z_{21} of two coupled antennas is defined as the negative of the emf V_{12} induced in antenna 2 by the current I_1 flowing in antenna 1 divided by I_1 with the terminals of antenna 2 open-circuited. By reciprocity $Z_{12} = Z_{21}$.

Electromagnetic Wave. A wave in which there are both electric and magnetic fields. Electromagnetic waves are known as radio waves, heat rays, light, X rays, etc., depending upon the frequency.

Transverse Electromagnetic Wave (TEM). An electromagnetic wave in which both the electric and magnetic fields are transverse to the direction of propagation.

Linearly Polarized Wave. A *TEM* wave in which the electric (or magnetic) field has a constant direction at a point in space.

Direction of Polarization. In a linearly polarized electromagnetic wave, the direction of the electric-field vector. Hence, the direction of polarization is vertical in a *vertically polarized wave,* horizontal in a *horizontally polarized wave,* tangent to the meridian line of a spherical frame of reference in a theta (θ) polarized wave, and tangent to the lines of latitude of the spherical reference frame in a phi (ϕ) polarized wave. The usual frame of reference has the polar axis vertical and the origin at, or near, the antenna.

Cross Polarization. The component of the electric-field vector normal to the desired polarization component.

Fraunhofer Region. That region of the field in which the energy flow from an antenna proceeds essentially as though coming from a point source located in the vicinity of the antenna. If the antenna has a well-defined aperture D in a given aspect, the Fraunhofer region in that aspect is commonly taken to exist at distances greater than $2D^2/\lambda$ from the aperture.

Fresnel Region. The region between the antenna and the Fraunhofer region.

Principal E Plane. A plane containing the direction of maximum radiation and in which the electric-field vector everywhere lies in the plane.

Principal H Plane. A plane containing the direction of maximum radiation and in which the electric-field vector is everywhere normal to the plane while the magnetic vector lies in the plane.

Power Density (Poynting Vector). In a given direction, the time rate of energy (*i.e.,* power) flow per unit area in that direction.

Radiation Intensity. In a given direction, the power radiated from an antenna per unit solid angle in that direction.

Radiation Pattern. A graphical representation of the radiation from an antenna as a function of direction. Cross sections in which radiation patterns are frequently given are the vertical and horizontal planes, or the principal electric and magnetic planes.

Radiation Lobe. A portion of the radiation pattern bounded by one or two cones of nulls.

Cone of Nulls. A conical surface formed by directions of negligible radiation.

Major Lobe (Main Beam). The radiation lobe containing the direction of maximum radiation.

Minor Lobe (Side Lobe). Any lobe except the major lobe.

Half-power Width of a Radiation Lobe. In a plane containing the direction of maximum of the lobe, the full angle between the two directions in that plane about the maximum in which the radiation intensity is one-half the maximum value of the lobe. *The tenth-power width* is the corresponding angle between the directions in which the radiation intensity is one-tenth (-10 db) the maximum of the lobe.

Directive Gain. In a given direction and relative to an isotropic antenna, it is 4π times the radiation intensity in that direction divided by the total power radiated by the antenna. Also the radiation intensity in that direction divided by the average radiation intensity radiated by the antenna (sometimes called gain; see *power gain*).

Directivity. The value of directive gain in the direction of its maximum value.

Power Gain. In a given direction and relative to an isotropic antenna, it is 4π times the radiation intensity in that direction divided by the total power delivered to the antenna (sometimes called gain; see *directive gain*).

Radiation Efficiency. It is the ratio of the power radiated to the total power supplied to the antenna at a given frequency (power gain = radiation efficiency × directive gain).

Gain of an Antenna (old usage). The measured gain of one transmitting or receiving antenna over another is the ratio of the signal power one produces at the receiver input terminals to that produced by the other, the transmitting power level remaining fixed.

Effective Area (Effective Aperture, Absorption Cross Section). In a given direction, the power in the terminating impedance divided by the power density of the wave incident upon the antenna. For a matched antenna, it is the square of the wavelength multiplied by the power gain (or directive gain) in that direction and divided by 4π.

Effective Height (present usage). The height of the antenna center of radiation above the effective ground level. For an antenna with symmetrical current distribution the center of radiation is the center of distribution. For an asymmetrical current distribution the center of radiation is the center of the current moments when viewed from directions near the direction of maximum radiation.

Effective Height (former usage). The induced voltage at the terminals of an antenna divided by the strength of the incident electric field. In l-f applications, as applied to loaded or unloaded vertical antennas, it is the actual height of the vertical section multiplied by the ratio of the average value of rms current in that section to the rms input current.

$$h = \frac{\int_0^L I(dl)}{I_{\text{in}}} = \frac{\text{meter-amperes}}{I_{\text{in}}}$$

Band Width. The range of frequencies within which the antenna performance, in respect to some characteristic, conforms to a specific standard. Impedance band width, for example, is often taken as the frequencies at which the input impedance departs from the resistance by a certain amount in terms of power factor; or in transmission-line coupled systems, the departure of the standing-wave ratio (SWR) from unity by a specified amount.

Antenna Tuning. Resonating an antenna system to some frequency other than a natural frequency by means of reactive devices.

Antenna Loading. The procedure where lumped reactances are connected in the antenna system for the purposes of antenna tuning.

Node. Point of zero or minimum voltage or current.

Antinode. Point of maximum voltage or current.

2. Radiation Control. Almost all modern radio applications require some specified form of radiation control for power gain, antifading, reduction of interference, radio guidance, reduction of multipath distortion, or reduction of noise. Antenna technology plays an important role in the general struggle to improve utilization of the radio spectrum.

Radiation is controlled by setting up current distributions, or aperture illuminations, which produce wave-interference patterns in space having specified or desirable forms best adapted to the wave-propagation circumstances. Generally, precise control in three dimensions is necessary. The higher the frequency, the easier it becomes to achieve high degrees of radiation control, limited only by frequencies so high that structures of conductors become too small to be practical and one must resort to

horns, reflectors, and other configurations. At the l-f end of the spectrum, the limitations are due primarily to great size (and cost).

The principles of wave interference have been highly developed in electromagnetic theory, and there is a vast literature with many excellent recent contributions.[1]

At a distance of several wavelengths from an antenna only the radiation field needs to be considered. This field is characterized by the fact that the electric and magnetic components are in time phase and are perpendicular to each other and to the direction of propagation. The strength or intensity of the field is inversely proportional to the distance from the antenna.

However, close to the antenna the field structure is quite complex because of the existence of the induction field of the antenna. The intensities of these fields decrease much more rapidly than the radiation field, and for an electric dipole, for example, the two are equal at a distance equal to the wavelength divided by 2π (approximately $\frac{1}{6}\lambda$). The induction field is important when antenna elements are coupled together by virtue of their proximity. Their existence must be kept in mind when making field-strength measurements near the antenna.

3. Current Distribution in Linear Conductors. The treatment of an antenna as an open-ended uniform transmission line with uniformly distributed attenuation is not in full accord with modern physical theories. However, this viewpoint is the simplest approach to a very complicated subject and is sufficiently in accord with practical results to justify its presentation in this form in a treatise of this type.

The various conductors of a radiating system carry currents of varying amplitudes, phases, and configurations in three dimensions. The fields set up by all the infinitesimal current elements of a system, together with the reflections which occur from the earth and reflecting areas, produce a characteristic radiation pattern. The problem in design is to set up current distributions which produce a desired radiation pattern. For this purpose, two basic types of distribution may be mentioned: those derived from standing waves on the conductors, and those derived from traveling waves. The first may be coarsely said to be (for a conductor length less than $\lambda/2$) a current distribution of constant phase and variable amplitude. The second is essentially a constant-amplitude and variable-phase distribution. It aids understanding of design principles to recognize these differences, even though practical systems have both characteristics in varying proportions.

In addition to their effect on the spatial distribution of radiant energy, current distributions determine the circuit impedance of an antenna, and the engineer must have regard for this fact for a number of reasons. Many forms of radiation patterns are realizable only by designing for predetermined impedance relationships at various points in the antenna and feeder system. As antenna techniques are developed, it becomes necessary at times to depart from simple natural current distributions and set up special distributions in linear conductors, sheet conductors, and loops. Antennas are really a complex of radiating portions and feeder portions, which have to be considered together. Only in the most elementary systems can the radiators and the feeders be treated separately.

4. Stationary-(Standing-)wave Distributions.[2] On a thin uniform linear conductor of arbitrary length with its far end open, excited at one end by an a-c generator, the current distribution can be considered as the result of interference between a wave traveling outward from the generator and another wave, due to reflection from the open end, traveling backward to the generator. The resultant current distribution is

$$I(x) \doteq \frac{V_R}{2Z_0}\left[e^{(\alpha+j\beta)x} - e^{-(\alpha+j\beta)x}\right] = \frac{V_R}{2Z_0}\sinh(\alpha + j\beta)x$$

where $I(x)$ = current amplitude and phase at a distance x from the open end of a conductor of length l including the apparent elongation Δx due to end effects

[1] See references listed under General Textbooks at the end of this chapter.
[2] See reference 28.

α = attenuation constant, nepers per unit length (1 neper = 1 hyperbolic radian = 8.686 db)
β = phase constant = $2\pi/\lambda$ radians per unit length (λ = wavelength measured on the conductor)
V_R = potential at the open end
Z_0 = characteristic impedance of the conductor $\doteq$ 138 $\log_{10}$ (2 × height ÷ diameter) for a vertical grounded radiator[1]

If α is negligibly small,

$$I(x) \doteq j \frac{V_R}{2Z_0} \sin \beta x$$

The physical significance of this is as follows. A current wave enters the conductor and describes a logarithmic spiral of decrement α and negative phase rotation β (determined by the velocity of propagation) as it passes along the wire, gets inverted at the end, and returns to the generator. The phasor (vector) sum of the incident and reflected current waves at any point x is $I(x)$.

In the theoretical case of $\alpha = 0$, the amplitude of $I(x)$ will vary sinusoidally with distance and remain constant in phase except for successive reversals at points which are multiples of $\lambda/2$ from the end. In reality, radiation and heat losses give finite values to α which are normally small. This causes $I(x)$ to depart slightly from a sinusoidal distribution, especially in the vicinity of the nodes, and causes the phase to change slowly with distance except in the vicinity of the nodes where it rapidly changes its phase to substantial reversal. To a good first approximation, then, the current distribution on thin uniform linear conductors can be assumed sinusoidal for purposes of calculating radiation patterns. A comparison between theoretical sinusoidal and measured current distributions on thin wires is shown in Fig. 1. G is the electrical length of the antenna expressed in degrees.

FIG. 1. Current distribution on vertical antennas of three lengths and small cross section. Circles are measured values; solid lines, true sinusoidal distribution.

In the same manner, the potential distribution along the conductor can be written

$$E(x) \doteq \tfrac{1}{2}E_R[e^{(\alpha+j\beta)x} + e^{-(\alpha+j\beta)x}] = \tfrac{1}{2}E_R \cosh (\alpha + j\beta)x$$

If $\alpha = 0$, the potential distribution is cosinusoidal and reverses its phase at values of x which are odd multiples of $\lambda/4$, and $E(x)$ is always in quadrature with $I(x)$ at any point. For small finite values of α, $E(x)$ departs slightly from this ideal amplitude and phase distribution, especially in the nodal regions, and $E(x)$ has a phase relationship with respect to $I(x)$ which is always less than 90 deg and comes into phase with $I(x)$ every quarter wavelength from the end.

Owing to the end effect, the distance to the first potential minimum is always

[1] See reference 12, p. 288.

slightly less than $\lambda/4$ from the open end. The amount varies from 2 per cent for thin wires to 10 per cent for large-diameter conductors and for the capacitive effect of insulators.

The impedance at any point in the system, looking toward the open end, is the vector ratio $Z(x) = E(x)/I(x)$. $Z(x) = R \pm jX$ is called the impedance as seen from that point (usually of interest only at a feed point).

Experimentally determined curves of resistance and reactance of cylindrical antennas normal to a perfectly conducting ground are shown in Figs. 2a and 2b.[1] For isolated balanced antennas the values shown must be doubled. These are shown in terms of the parameter D, which is the diameter of the antenna in electrical degrees and includes the capacitance of the closed lower end of the cylinder which in all cases is 1 deg from ground. These curves have fundamental reference value.

FIG. 2a. Measured resistance of cylindrical antenna normal to perfect ground.

Empirical Distributions. By modifying the velocity of propagation and the characteristic impedance of the conductor, uniformly or nonuniformly, or by use of lumped reactances inserted in the system, new expressions for the current and potential distributions are required. Empirical distributions can be produced in this way.

5. Traveling-(Progressive-)wave Distributions. If instead of being open the end of this conductor is terminated so as to absorb all energy arriving at the end without reflecting any of it, then a traveling-wave system results. Under perfect realization of this condition $Z(x) = Z_0$ is constant for any value of x.

In the standing-wave condition, impedance variations with length vs. frequency exhibit properties analogous to resonance (at x = odd multiples of $\lambda/4$) or antiresonance (at x = even multiples of $\lambda/4$). These resonance effects are selective to frequency in varying degrees and limit the bandwidths that can be obtained. In the traveling-wave condition, selectivity is impossible unless the termination is selective.

6. Special Distributions. Arrays of conductors having predetermined current distributions, amplitudes, and relative phases of excitation produce special distributions of current in three dimensions for radiation control. By so doing, each conductor is immersed in the fields of all the others, producing energy interchanges through radiation couplings. They produce circuital effects called mutual impedances. Depending on spacing, relative current amplitudes and phases, and orientations, the mutual

[1] See reference 30.

impedance between two radiators at a given frequency is in general complex and can be vectors lying in all four quadrants. R and X can both be either positive or negative.[1]

7. Distributed Potentials Induced by Space-wave Fields. When one conductor is in the field of another, the energy exchange is by means of potentials induced in all parts of the wire. These potentials would have to be simulated by a system of distributed generators in contrast with a single generator used to excite a single radiator. Distributed generators produce a different current and potential distribution along a wire than does a single generator such as a transmitter. Therefore, mutual-impedance effects modify the shape of the current distributions from those prevailing in the absence of mutual impedance.

FIG. 2b. Measured reactance of cylindrical antenna normal to perfect ground.

In the case of receiving antennas, which are excited entirely by the field of the passing electromagnetic wave, the current and potential distributions are not like those for transmitting except for a $\lambda/4$ grounded antenna and a $\lambda/2$ antenna above ground, as shown in Fig. 3.

V-L-F AND L-F ANTENNAS (10 TO 300 KC)

With the advent of high frequencies for point-to-point communications, the low frequencies were relegated to the past in the popular mind. There has been a revival of the use of low frequencies for transmission in the auroral zones where the high frequencies are too unreliable for certain important services such as aviation. An exception to this is the recently discovered ionospheric forward-scatter mode of propagation in the v-h-f band which is unaffected by auroral disturbances. Fortunately the auroral zones are among the lowest-noise zones of the world so that, aside from the factor of propagation, the reliability and the effectiveness of low frequencies are much greater than for the high-noise zones of the lower latitudes.

The efficiency and band width of antennas whose dimensions are much less than the wavelength are essentially limited by the antenna size. Small antennas, whose maxi-

[1] See references 28, 32, 49, 56, 58, 59, 67, 68, 76, 83, 119, and 124.

mum dimensions are less than $\lambda/2\pi$, can be made to behave essentially as lumped capacitances (electric dipoles) or inductances (magnetic dipoles or loops).[1]

The power available from a small antenna can be expressed in terms of its effective area, which is $3\lambda^2/4$ sq m. However, in each instance the unavoidable loss in the coupling circuits restricts the output to only a small fraction of the power input. While the radiation pattern and hence directive gain of a small antenna remain the same as its size is decreased, the radiation resistance decreases relative to the other resistance in the coupling circuit. The resulting reduction in coupling efficiency is one of the principal limitations of small antennas. The reduction in size also imposes a fundamental limitation on band width, since a smaller antenna must be more sharply tuned to deliver its available power. Increased bandwidth can be obtained only at the expense of coupling efficiency.

V-l-f and l-f antennas are generally characterized by short electrical length with consequent low radiation resistance and high input reactance. Design techniques feature methods for maximizing radiation resistance and minimizing ground resistance, insulation loss, and reactance. Power input is limited by corona or flashover potentials. The potentials on the system are uniform within a few per cent. Useful radiation is that which is vertically polarized, and service is rendered by ground wave with some D-layer influence. Ground-system design must take into account the need to minimize current densities at collecting points and the very deep penetration of ground currents into soils of highest conductivities and even into sea water. This is the oldest branch of antenna technology, and yet there is room for additional research.

FIG. 3. Current and charge distribution in thin uniform wires for conditions of free oscillation. In the receiving case, the antenna is unloaded and oriented parallel to the electric field.

8. Fundamental Wavelength. Consider the vertical top-loaded antenna in Fig. 4. Because of its greater size at the top, there results a modified current distribution which produces a more desirable input impedance and vertical plane radiation pattern. In this figure, V = vertical length (height) of the down lead, H = horizontal length of top-loading section. D = diameter of vertical portion. Note that at any frequency

[1] See reference 99.

an antenna dimension can be expressed in wavelengths by the ratio of the physical length to the wavelength (expressed in same units as length) corresponding to the frequency, or in electrical degrees (or radians) by the product of the dimension in wavelengths and 360 deg (or 2π radians). For example, at a frequency of 10 kc, *i.e.*, 3×10^4 m wavelength, the vertical length of an antenna 300 m high can be expressed as $V = 300$ m $= 0.01\lambda = 3.6$ deg $= 0.0628$ radian.

The lowest frequency at which the antenna system in Fig. 4 is resonant is defined as the fundamental frequency, and the corresponding free-space wavelength is defined as the fundamental wavelength λ_0. The effective (or apparent) electrical length G of this antenna system at the fundamental frequency is $\lambda_0/4$, or 90 electrical degrees. For any vertical length V, the effective length is a function of the length of the top-loading section H and the diameter of the vertical section D.

FIG. 4. Empirical values for the constant K used in determining the wavelength of T antennas.

If the fundamental frequency, and hence wavelength, of this antenna system is known, its effective electrical length G at any other frequency can be obtained from

$$G = \tfrac{1}{4}(\lambda_0/\lambda) \quad \text{wavelengths}$$
$$= 90(\lambda_0/\lambda) \quad \text{degrees}$$
$$= (\pi/2)(\lambda_0/\lambda) \quad \text{radians}$$

where λ is the wavelength corresponding to the frequency.

The fundamental wavelength can be obtained from the following approximate relations or from measurements on scale models. For an antenna without top loading, or a vertical tower of uniform cross section, the following formula applies to within approximately 4 per cent.

$$\lambda_0 \doteq 1.33V \quad \text{m}$$

where V = vertical length (height), ft.

For top-loaded symmetrical T antennas the fundamental wavelength can be predicted only approximately from simple formulas, and for large-area flattop systems it usually must be estimated or measured from scale models. Many empirical factors are the cause of this uncertainty, including the ratio of horizontal to vertical lengths, the capacitance of insulators, the proximity of grounded or insulated towers, the number of conductors used and their cross-sectional configuration, reflections from the T junction, and perhaps other factors. It has been found from practical experience that the following empirical relation for the fundamental wavelength can be used for a symmetrical T antenna.

$$\lambda_0 \doteq K(V + H/2) \quad \text{m}$$

where the antenna dimensions are in feet, λ_0 in meters, and K is taken from Fig. 4. This figure was derived from 24 actual antennas which were carefully measured. The antennas were of different sizes and with several different tower arrangements. This unusual curve accounts within 5 per cent for all but four of the antennas. The corners in the curve are based on reliable data and so were not rounded out. The value at the right-hand extremity of the curve is based on three identical l-f antennas constructed in three different locations and measured by separate qualified individuals with exact concordance.

In some cases the effective electrical length G of an inverted L or T antenna can be taken as the sum of the vertical and horizontal parts in electrical degrees (or radians) provided that the horizontal distances from the junction with the vertical are electrically small. For capacitance-area loading of empirical form, calculations may be difficult. In such cases, it is desirable to use a scale model (the scale being in size

directly proportional to the wavelength) erected over a metallic plane of area very large with respect to the flattop. The supports should also be modeled and the location of antenna and guy insulators simulated. For reactance measurements, the cross section of the antenna conductors should also be to scale.

9. Radiation Resistance and Field Strength. The radiation resistance can be increased by increasing the vertical length (see Fig. 5) and/or by producing, as nearly

Fig. 5. (*Left*) Radiation resistance for electrically short antennas vs. height of vertical portion (degrees) and (*right*) ratio of currents at top and base of vertical portion.

as possible, a uniform current distribution on the vertical portion of the antenna. The latter is usually done by some capacitance loading at the upper end in the form of an inverted L, a T, or a large flattop in the form of a ring or spider.

The radiation resistance of a loaded antenna can be calculated from the following (see Fig. 5).

$$R_r = \frac{30}{\sin^2 (L + B)} \left[\frac{4L^2}{3} \sin^2 B + \frac{2L^3}{3} \sin 2B + \left(\frac{17}{30} \cos 2B - \frac{7}{30} \right) L^4 \right] \quad \text{ohms}$$

where L = vertical length (height), deg

B = effective lengthening of the antenna due to the top loading, deg; $L + B = G$

When $B = 0$, $R_r = 10L^2$ (L in radians). The above relationships are valid for $L \leqq \lambda/12$ ($L \leqq 30$ deg) and are shown in Fig. 5.

If one assumes a linear current distribution on the antenna, as shown in Fig. 5, then

$$I_{top}/I_{base} = B/B + L = 1 - L/G$$

where L, B, and G are expressed in degrees.

It is a great convenience to determine l-f antenna performance in terms of meter-amperes in the vertical portion (see definition of effective height). In most l-f antennas, the current distribution on the downlead is linear, and the ratio of current near the flattop to that near the base can be accurately calculated from the electrical dimensions, as indicated above, permitting in turn an accurate calculation of the meter-amperes provided the current can be observed at the lower end of the downlead,

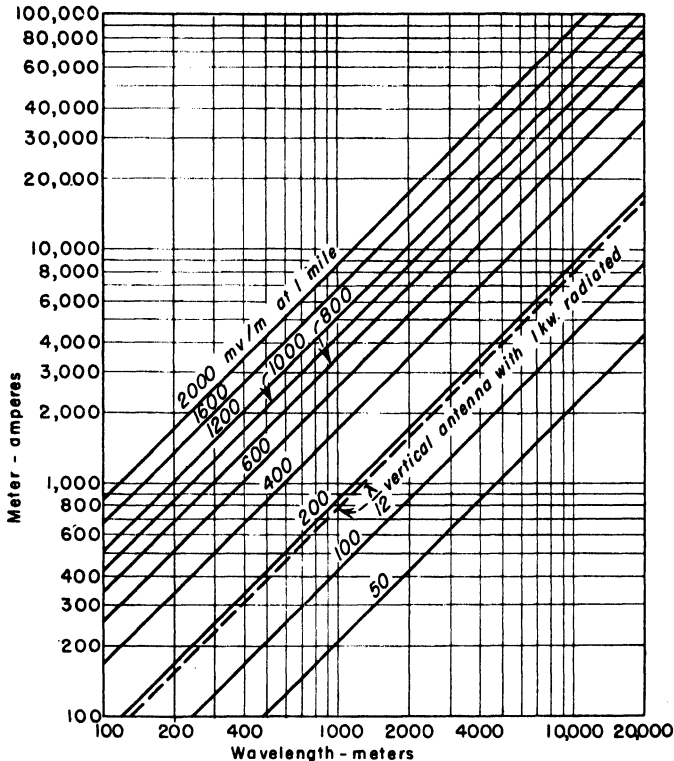



FIG. 6. Meter-amperes required to produce given field at a given wavelength.

above all tuning apparatus. Field intensity is directly proportional to meter-amperes at any one frequency. Since a perfectly efficient λ/12 vertical antenna with 1,000 watts radiated power gives a field of 187 mv per m at 1 mile, an excellent reference for determining radiated power is available. (This reference is used because longer vertical antennas begin to provide some gain by vertical directivity.) Then if the total antenna resistance can be measured and the total power input determined, the radiation efficiency can be computed from field strengths (see Fig. 6).

The maximum field intensity possible with a given power input is the objective of all antenna designing. With l-f systems the cost per decibel of improvement at times may be very great, especially at the lowest frequencies.

The measurement of antenna resistance of antennas of very short electrical length is a difficult operation and requires special power oscillators to permit the use of thermocouple

ammeters having internal resistances small with respect to that of the circuit being measured. The resistance-variation method is also difficult unless special heavy-duty standard-value resistors are provided. The reactance-variation method is difficult because of the range of variability required in a heavy-duty tuning device. The r-f impedance bridge is an excellent instrument for such measurements.

The most convenient method of measuring antenna performance is by field-intensity measurements at a standard distance. Since large antennas which are multiple-tuned are not point-source radiators, substantial distances are required for observing performance of the v-l-f systems. For antennas with single tuning, distances as close as one wavelength may be used for measuring the radiation field. For wavelengths shorter than 1,600 m, 1 mile is a desirable standard distance, measured from the downlead.

10. Scale-model Measurements.[1] Valuable data on an l-f antenna can be obtained quickly and economically by h-f scale-model measurements. Two very important values, reactance and fundamental wavelength, can be determined accurately using laboratory instruments such as a Q meter or an r-f impedance-measuring device.

Resistances of ground, conductors, and insulators cannot be simulated. It is, therefore, desirable to build the scale model over a large continuous metallic plate acting as a perfect ground. The linear and cross-sectional dimensions are to the same scale. Towers and guys can be simulated precisely, and the effects of placement of insulators in supports and guys can be measured.

The fundamental wavelength can be measured by connecting the downlead to the ground plate and shunt-exciting the system with very loose coupling from a buzzer. The latter must be shielded. The emitted wavelength is measured by reception using a calibrated receiver. Also any of the familiar techniques of measurement of resonance in the antenna with a coupled oscillator may be used if preferred. The buzzer method is very simple and satisfactory.

Reactance can usually be measured over the range of values of interest using a Q meter. For conditions of high Q values, body effects near the antenna produce errors which can be eliminated by manipulating the Q meter from a distance. Since capacitive reactances are being measured, appropriate methods must be employed.

Scales as extreme as $\frac{1}{100}$ of natural size have proved successful with careful measuring techniques. Scales of the order of $\frac{1}{25}$ are more easily measured, but the model may be very large. The ease with which a number of empirical constructional details can be measured and compared permits the final design to be quite thoroughly evolved in the model stage. The potentials, insulator arrangement, conductor arrangements, rigging details, support locations, tower heights, etc., can be developed as far as desired.

The principle of similitude upon which model tests are based can also apply in the opposite direction by using, for instance, v-h-f models for microwave antennas, where the enlarged antenna may be more easily measured and developed than at natural size. In all scale-model work, the frequency used is proportional to the scale of the model. At $\frac{1}{50}$ scale, 50 kc is simulated by 2.5 Mc.

11. Reactance. The reactance of the antenna at the operating wavelength λ is an important value to determine so as to know the required size of the antenna load coil (l-f antennas always have a negative reactance) and the antenna potentials encountered. The reactance must be minimized to reduce loading inductance (and load-coil resistance) and to reduce insulation loss even when potentials are far below that which produces corona. This also reduces the selectivity of the antenna system, permitting higher keying speeds without excessive telegraph distortion, or greater frequency range for telephony. Reactance is reduced by increasing the top-loading capacitance, increasing the cross section of the antenna with cage or flat multiple conductors, or both. As a first approximation, the antenna can be considered as a lossless transmission line, since the reactive component of the input impedance is usually very much greater than the antenna resistance. In this case, the input reactance is

$$X \doteq -jZ_0 \cot G \qquad \text{ohms}$$

[1] See reference 28.

where G = effective electrical length of antenna, deg

$$Z_0 = \sqrt{L/C} \doteq 138 \log_{10} \frac{2 \text{ length}}{\text{diam}}$$

L = inductance per unit length
C = capacitance per unit length

These relations are satisfactory for $G \leqq 30$ deg.

It is seen that cot G is important but is determined by the electrical size of the antenna and has a large influence on its cost. For a given length and size of structure, cot G is constant. Further reduction of reactance must therefore be made by reduction of Z_0, which is done by decreasing inductance and increasing capacitance per unit length. There are mechanical limitations to cages or other forms of multiwire conductors because of wind and sleet, the increased sags and tower loadings, and the

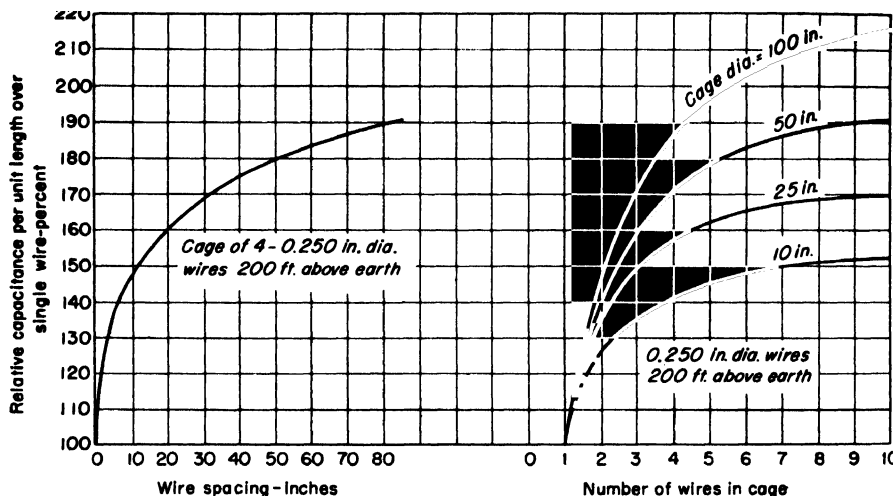



FIG. 7. Increase in capacitance of cage vs. single wire.

multiplied problems of vibrational stress and fatigue. Figure 7 shows the increase in capacitance per unit length that was obtained for a top-loading cage in terms of percentage increase over a single wire and illustrates in general how the capacitance per unit length of a vertical l-f antenna can be increased by increasing the cross section. Of course, there is a point where mechanical objections outweigh electrical advantages.

12. Antenna Potential. The potential V_a at the feed point of a vertical l-f antenna (near ground) for a power input W into the antenna impedance $Z_a = R_a - jX_a$ is

$$V_a = I_a Z_a$$

The antenna current $I_a = \sqrt{W/R_a}$.

Most l-f antennas (or any electrically short antenna) have a reactance large with respect to the resistance, so that usually

$$|V_a| \doteq I_a X_a$$

The antenna potential is kept to a minimum for practical reasons of simplifying the insulation and avoiding corona or flashover. Therefore, the reactance of the antenna is made as low as practicable for a given total antenna resistance and power input.

There is a rise in potential along the antenna above the feed point, the amount varying with the electrical length of the system and depending upon the configuration

of the conductors. For a uniform vertical antenna, the maximum potential becomes

$$V_{max} \doteq \frac{V_a}{\cos G}$$

In the case of an antenna with a very-high-capacitance flattop,

$$V_{max} \doteq \frac{V_a}{\cos L}$$

where L = vertical length (height) of antenna, deg.

Prevention of corona, pluming, and flashover is a matter of both conductor and insulator design. Potential gradients must be kept below critical values for the ionization of air at the altitude of the site, under all meteorological conditions. Moisture films on insulators and drip water on conductors often increase losses or cause corona or plume formation, especially in still air. Adequate allowance has to be made for insulation in any case, but when antenna potentials surpass 100 kv, very special design problems are involved even at sea level; if altitudes of 5,000 to 10,000 ft are encountered at the site, the power may be severely limited, or the fundamental design of the system affected. In the latter case, it is simpler to employ self-supporting towers as radiators to attain low gradients for relatively high potentials by virtue of the large cross section of the tower. This centers the flashover problem in the base insulation, where it is not too difficult to take the required precautions for satisfactory design.

13. Antenna Resistance. The electrical design of an l-f antenna must maximize radiation resistance and minimize conductor, insulation, and ground components of resistance. In the antenna circuit is also the tuning inductance, the resistance of which must also be minimized. In a well-designed system, ground and tuning-coil resistances are the principal sources of serious power loss, and special attention must be directed to these matters to obtain maximum radiation efficiency.

14. Multiple Tuning.[1] V-l-f antennas often have radiation resistances which are only a fraction of an ohm. To obtain a radiation efficiency even as high as 6 per cent it has been necessary to use several downleads from the horizontal section and tune each downlead separately. The antenna is fed from one point only. This technique of multiple tuning was an outstanding development in the improvement of v-l-f antenna efficiency. The use of a separate ground system under each downlead reduces ground-current densities and, therefore, total ground resistance, because the total antenna current is divided among the several ground systems. The radiation resistance as seen from one downlead increases as n^2, where n is the number of multiple-tuning downleads.

15. Ground Systems. The reduction of ground resistance starts with the selection of the site, which should be an area of highest available conductivity to a considerable depth and for a distance of more than $\lambda/2$ from the antenna. The site need not be topographically flat although an undulating surface complicates construction. Surrounding hills or mountains should have an altitude which is a small part of a wavelength to avoid shadows and high attenuation for ground-wave propagation. From the standpoint of ground-system resistance alone, however, conditions beyond $\lambda/2$ from the antenna are relatively unimportant.

The depth of the soil of good conductivity is important because of the very deep penetration of earth currents at low frequencies. In general, however, the earth is a relatively poor conductor and provision must be made for returning these currents to the feed point at the base of the antenna. A large radial ground system buried in the ground surface will collect currents due to the field in space over it and will guide them back to the feed point. It is often economically impractical to use ground systems of the radial type longer than about 500 m, and this may be but a small portion of a wavelength. The fields beyond the wires set up ground currents which return to the feed point through the soil. Therefore, in this case a considerable portion of the

[1] See references 103, 104, and 110.

ground current returns to the feed point at the antenna base under the ground system, tending to produce high current densities at that point. If long ground rods are attached to the ends of the wires, these will intercept some of the ground currents near the surface where they have the greatest magnitude and refract some of the current into the wires from underneath, but the deeper currents will still pass under the system and the greater portion will be concentrated in the soil as they approach the feed point. Therefore, special means must be taken to collect these currents returning from a great depth in a manner to minimize current densities in the soil and thus reduce ground-system resistance.[1] A circle of long ground rods attached to the ground wires a short

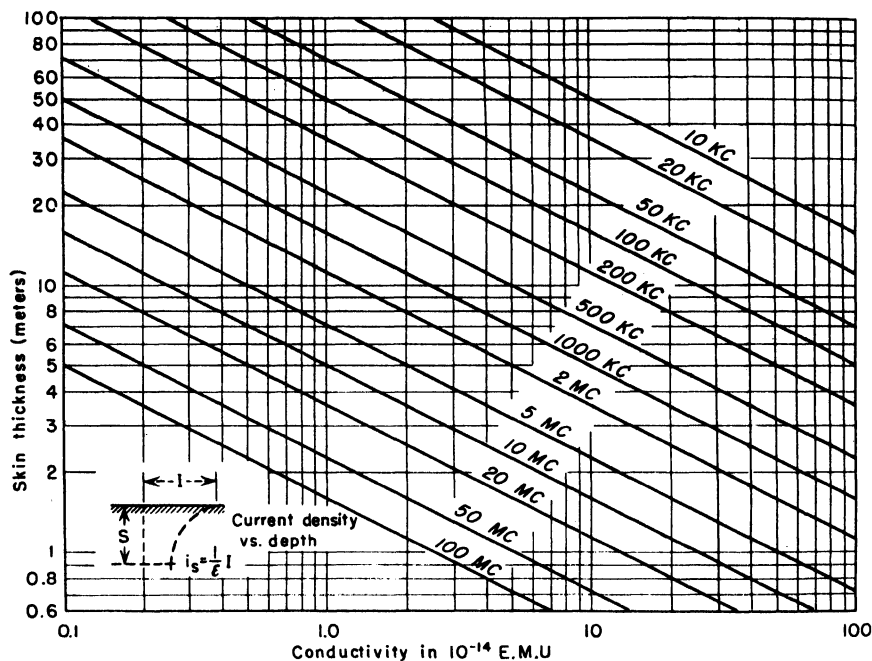



FIG. 8. Penetration of ground current. Skin thickness $S = (4\pi^2 \alpha_{emu} f_{kc} \times 10^3)^{-1/2}$ cm.

distance from the feed point is desirable, as is also a large pipe driven deep into the soil directly under it and connected into the ground system. The length and number of radial wires are quite empirical, depending upon frequency, soil characteristics, and economics.

Another method of ground-system design for v-l-f systems is to employ from one to several circles of four to six short star ground meshes, each of six to twelve 100-ft radial wires with long ground rods at their ends. From the center of each star a conductor is brought back above ground, or insulated from ground, to the feed point or multiple-tuning point. When more than one circle of such stars is used, ground-equalizer inductors are inserted in the shorter return wires so as to equalize the currents collected from each star. This technique reduces current densities in the soil by collecting ground currents in equal amounts from a large symmetrical area surrounding the antenna downlead.

Figure 8 shows depths of penetration of earth currents for various soil conductivities and frequencies. Approximately 90 per cent of the power loss in the soil occurs within

[1] See reference 121.

the so-called skin thickness, which is that depth at which the current density has diminished to 37 per cent of that at the surface.

Capacitance grounds, or counterpoises, have certain merits which have long been recognized.[1] Instead of collecting ground currents by conduction from the soil, a large capacitance area composed of a system of wires a few feet above the earth is used. A virtually uniform electric-field distribution exists under the counterpoise and causes an equally uniform distribution of earth currents beneath. The fringe effect increases its effective area a few per cent. The area covered by a counterpoise should be as large as practicable to minimize ground losses. A small counterpoise located near the base of the downlead or tower, sometimes called a ground screen, may in some cases reduce local ground losses when used in conjunction with a buried-wire system.

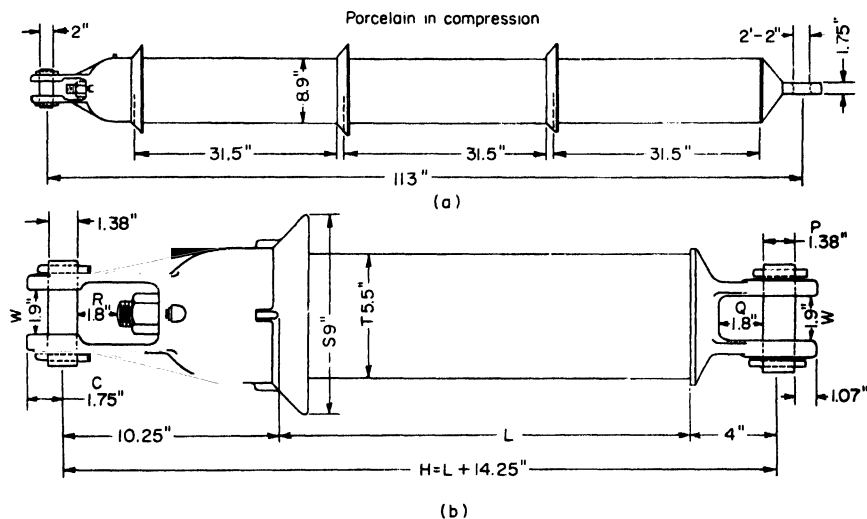



FIG. 9. (a) Heavy-duty strain insulator, safety-core type, preloaded, oil-filled. (b) Safety-core strain insulator, preloaded link-core type.

16. Antennas for L-f Broadcasting. In those countries using the frequencies from 160 to 265 kc for broadcasting, there has been economic justification in employing antennas operating at or near the fundamental wavelength ($G = 90$ deg). This results in a relatively high radiation resistance and low reactance. In Europe there has been rather general adoption of a T-type single-wire antenna supported between two very high towers (as high as 250 m in some cases).

Droitwich Antenna.[2] A typical example which has been thoroughly described in the literature is that of the BBC station at Droitwich, England. A T antenna with a flattop 550 ft long and a downlead of 630 ft is supported between two towers 700 ft high and spaced 800 ft. The antenna consists of one wire. At the operating frequency of 200 kc, the impedance is $21 + j0$; at 190 kc, it is $17.5 - j41$; at 210 kc, it is $24 + j41$ ohms.

Sometimes, to correct for antenna selectivity, antenna-coupling networks are applied which equalize the antenna impedance over a range of 4 to 6 kc each side of the carrier frequency. High single vertical radiators of large cross section, derived from the tower itself or by employing supported vertical wires around the tower to effect a large cross section, have the characteristics of lower intrinsic selectivity and high radiation efficiency.

17. High-tension Insulation for Wire Antennas. Very serious electromechanical problems are encountered in the insulation of large l-f antenna systems which must

[1] See reference 108.
[2] See reference 105.

operate at high power inputs. Long spans, large conductors, heavy downleads, large total projected areas exposed to wind and sleet, and minimized sags so as to keep effective heights at a maximum all contribute to enormous stresses at the insulator locations. Maximum electrical duty occurs simultaneously with maximum mechanical duty. Insulators must be of lowest possible capacitance, and the number used in parallel must be a minimum. Typical insulators with high mechanical strength for use with very-high-tension antennas are shown in Fig. 9.

18. L-f Radio Range Antennas (200 to 400 Kc). The use of low frequencies for radio guidance in aviation on a large scale in the form of four-course radio ranges, the German Zonne system, and radiophares generally has led to the vertical tower radiator for the elimination of night effect. Any emission of horizontally polarized energy from such systems causes large errors at night in courses and in direction finding. The Adcock type of radio range is now standard in the United States and Canada, uses four radiators (square array) or five (quincunx array) approximately 125 ft high spaced about 600 ft on a diagonal. These are usually base-insulated self-supporting towers.

Fig. 10*a*. Measured resistance of various whip antennas mounted on an 8-cwt truck with truck chassis as ground: (1) 8-ft single rod; (2) 12-ft single rod; (3) 16-ft single rod; (4) 8-ft V aerial, 60 deg on common base insulator; (5) 12-ft V aerial, 60 deg on common base insulator; (6) 16-ft V aerial, 60 deg on common base insulator.

There are several evolutionary versions of radiators and ground systems for the four-course ranges. With a ground system of fifteen 100-ft radial wires (buried) with end ground rods, for each radiator, the radiation efficiency is generally of the order of 8 to 10 per cent. Ground screens in the form of elevated metallic counterpoises about 30 ft square under the tower-base insulators are also used. They are series-tuned, and two outer radiators on a diagonal operate together to produce a figure-of-8 field-intensity pattern keyed systematically with the Morse A interlocked with N keying on the opposite diagonal pair. As power is switched between the two pairs of diagonal radiators, signals in the overlap region where the field intensities are equal produce a constant uninterrupted tone which provides a fixed track for guidance of aircraft. The ability of the ear to distinguish volume levels from the two alternated patterns gives a course width of about 3 deg. Just at the place where the A or N keying becomes observable each side of the course is a narrow angle called the *twilight zone*, and aircraft ordinarily fly in the right-hand twilight zone. Between courses the A or N signals predominate.

When equal power is fed to the two halves of this system, reciprocal 90-deg courses are set up. If the power is lower in one pair, its radiation pattern is smaller than the other and reciprocal courses result but the courses are said to be *squeezed.* When both power and phase differences in the radiator currents are other than 180 deg in each pair, nonreciprocal courses are obtained. This is called *course bending.* These combinations are used to set range courses at angles best suited to airway routings in the four directions. The antenna system is energized through a goniometer, the rotation of which will turn the entire radiation pattern to the orientation desired.

The four-tower nonsimultaneous range switches the four towers into parallel in-phase feed when voice broadcasting is required, and during this interval the range courses are eliminated. The simultaneous range using five towers permits the four outer ones to transmit range courses continuously for guidance while the center tower radiates voice signals when required.

19. Whip Antennas. Whip antennas used on vehicles for transmission of h-f waves present much the same basic design problem as the electrically short l-f antenna. Radiation efficiency is intrinsically low, and care must be taken that it is no lower than necessary. In such systems tuning-coil loss and ground loss consume almost all of the transmitter power, and radiation efficiencies of the order of 0.5 to 2 per cent are common in the 2- to 8-Mc band. Base-insulator leakage and capacitance should be as low as practicable. Anything that conditions of service permit which will decrease the antenna reactance will contribute much to over-all improvement in efficiency. Some typical whip-antenna data are shown in Fig. 10*a* and *b*.

FIG. 10*b*. Measured reactance of whip antennas: (1) 8-ft single rod; (2) 12-ft single rod; (3) 16-ft single rod; (4) 12-ft V aerial, 60 deg on common base insulator; (5) 16-ft V aerial, 60 deg on common base insulator.

20. Beverage (Wave) Antenna.[1] This type of antenna, one of the most effective l-f directive receiving systems to be used commercially, is a long transmission line. It is named after its inventor, H. H. Beverage, but is also called the *wave antenna.* In its very simplest form, the Beverage antenna is a single straight horizontal wire a few feet above grade level, the length being anywhere from one to several wavelengths.

[1] See references 106 and 107.

FIG. 11. Houlton antenna system for receiving transatlantic long-wave signals.

The characteristic impedance of this wire unbalanced to ground is roughly calculable by using the image in the ground as the second conductor in a parallel-wire system. The receiver is coupled in at one end of the line, and the other end is terminated in a resistance equal to the characteristic impedance. Stable ground systems are necessary at both ends.

The long open-wire transmission line pointed in the direction of a passing wave has a high degree of exposure to the horizontal component of the wave front because of the wave tilt that earth losses produce in vertically polarized low-frequency waves traveling along the surface of the earth. This induces in the line a continuous series of emfs that are propagated along the wires in the form of a traveling wave. That is, a passing wave sets up a traveling wave in the wire which starts at the distant end (in the direction of arrival of the passing wave) and is propagated toward the end where the receiver is situated. In this manner the entire wire receives energy from the passing wave so the effects are cumulative at the receiver. Energy collected from a passing wave traveling in the opposite direction is completely dissipated in the terminating resistor and so does not enter the receiver. Waves arriving from the sides have relatively little effect on the receiver. Hence, this antenna has high directivity in the horizontal plane.

This antenna is suitable only for reception since ground losses are relatively high compared with the radiated energy in transmitting.

The Beverage antenna has many useful forms which are specially adapted to long-wave reception, to short-wave reception, to bidirectional and unidirectional selectivity, for vertical and horizontal polarization, etc. A thorough treatment of these is impossible here, and detailed data must be obtained from the original and subsequent papers on the subject.

Directivity may be sharpened by using two or more antennas in an array. This has been done in the Houlton system, shown in Fig. 11, which is used for transatlantic telephone reception on long waves. One of the several forms of antenna used in this application is that which couples the receiver to the end of the antenna that is nearest the transmitting station.

A two-wire line is used to achieve this in the following manner. Waves arriving from the preferred direction act upon the two wires in parallel to ground, and the induced wave of energy in the wire travels to the distant end where it encounters a reactive network called a reflection transformer. This device reverses the phase of the wave in one of the wires and reflects the energy from the end back to the receiver, the reflected wave of energy now traveling in the two wires balanced to ground.[1] The receiver coupling network terminates the line and absorbs all the wave energy in actuating the receiver. A wave entering the system from the reverse direction travels along the two wires in parallel against ground, produces no potential difference across the balanced termination, and therefore has no influence on the receiver. Instead, the circuit to ground is terminated in the characteristic impedance of the parallel-grounded system, and the unwanted wave is completely dissipated in a resistor.

MEDIUM-FREQUENCY ANTENNAS (300 TO 3,000 KC)

At the lower frequencies, below 3 Mc, where ground-wave absorption is not too excessive, or at frequencies below about 5 Mc, where a low angle of radiation is required, the grounded vertical antenna is extensively used with vertical heights that range from only a few tenths of a wavelength to over one-half wavelength. This type of antenna is resonant when its effective length (height) is one-quarter wavelength. In general, the radiation is vertically polarized in an omnidirectional pattern in the horizontal plane.

The radiation-pattern characteristics of the grounded vertical antenna should be such as to confine its radiation to angles sufficiently close to the horizon that the nearest point to the transmitter at which the sky wave returns to earth is just beyond the limits of the ground wave to prevent severe fading. A vertical $\lambda/2$ structure (unloaded) is ideal but sometimes impractical because of its height.

[1] A similar effect is obtained by grounding one wire and leaving the other open-circuited.

The efficiency of these antennas depends upon the ratio of the radiation resistance to the total resistance of the antenna system, which includes loss resistance in the antenna, coupling system, and ground system; the latter losses are quite serious for antennas shorter than $\lambda/4$. From an efficiency point of view, the vertical $\lambda/2$ antenna is the optimum because the radiation resistance is very much greater than the loss resistances.

The shorter vertical antennas are generally brought into resonance by top (capacitive) loading or base (inductive) loading. Top loading is highly desirable, as in l-f application, because this increases the radiation resistance.

The prime interest in the m-f band has been, and still is, in the standard broadcast band (535 to 1,605 kc) because of its commercial aspects. Each station is assigned a carrier frequency, a 10-kc channel (the frequency band occupied by the carrier and two side-band frequencies with carrier in center), beginning at 540 kc.

When broadcasting began, the antennas used were smaller versions of the types previously used for lower frequencies. When the possibilities for getting higher field strengths along the ground by using high vertical radiators became known,[1] and greater knowledge of wave propagation in this band was obtained, a period of antenna development began which led to definite optimum designs,[2] which are physically and economically realizable. While studies of antenna structures were being made, other detailed study was directed to ground systems, and again some definite practical optimum designs resulted.[3] This work placed broadcast-antenna engineering on a firm foundation, which, together with the vast information on wave propagation by ground-wave and ionosphere reflections, has permitted an intense utilization of this band of frequencies for broadcasting services. This paved the way for the use of directive antenna systems which further increased the utilization of the limited number of channels for an ever larger number of stations and increased power.

Because of a number of fortunate natural circumstances, broadcast-antenna engineering is far simpler than that for most other frequency bands. For instance, a substantially perfect ground can be built under the antenna, permitting high radiation efficiency. Within realizable heights of structures, the vertical radiation pattern can be adjusted for maximizing the ground wave and minimizing the sky wave, thereby decreasing the annular zone of overlap of the two and giving a high degree of effectiveness to both types of propagation by concentration of radiated power at angles close to the horizon. With modern data, broadcast-antenna performance can be predicted and measured with a remarkable degree of precision.

21. Nondirective Broadcast Antennas. Vertical polarization is required for broadcasting in this band, and vertical radiators are required to produce it. One steel tower used as a radiator costs less than two towers supporting an equivalent wire antenna and eliminates field distortion caused by currents induced into the supporting towers. Hence present-day broadcast antennas are almost exclusively the tower-radiator type, either guyed or self-supporting. Structures having uniform or nearly uniform cross section are necessary to assure a predictable vertical radiation pattern from a predictable current distribution. Top loading, sectionalizing, and combinations of these two techniques are occasionally employed to produce an optimum vertical directivity with limited tower heights. Tower radiators have further advantages derived from their relatively large cross section, which provides a lower characteristic impedance than a wire antenna, with lower potential gradients for the same power, lower reactance, and greater band-width capability for a given impedance variation.

22. Current Distributions and Vertical Radiation Patterns. As the electrical length of a uniform vertical conductor is increased, a larger and larger portion of a complete near-sine wave of current is distributed on it. In space, each element of length produces a field whose intensity is directly proportional to its current amplitude. The integrated fields from all parts of the antenna, together with those reflected from the ground, interfere in varying degrees from complete addition to complete cancella-

[1] See references 19 and 20.
[2] See references 118, 120, and 139.
[3] See reference 121.

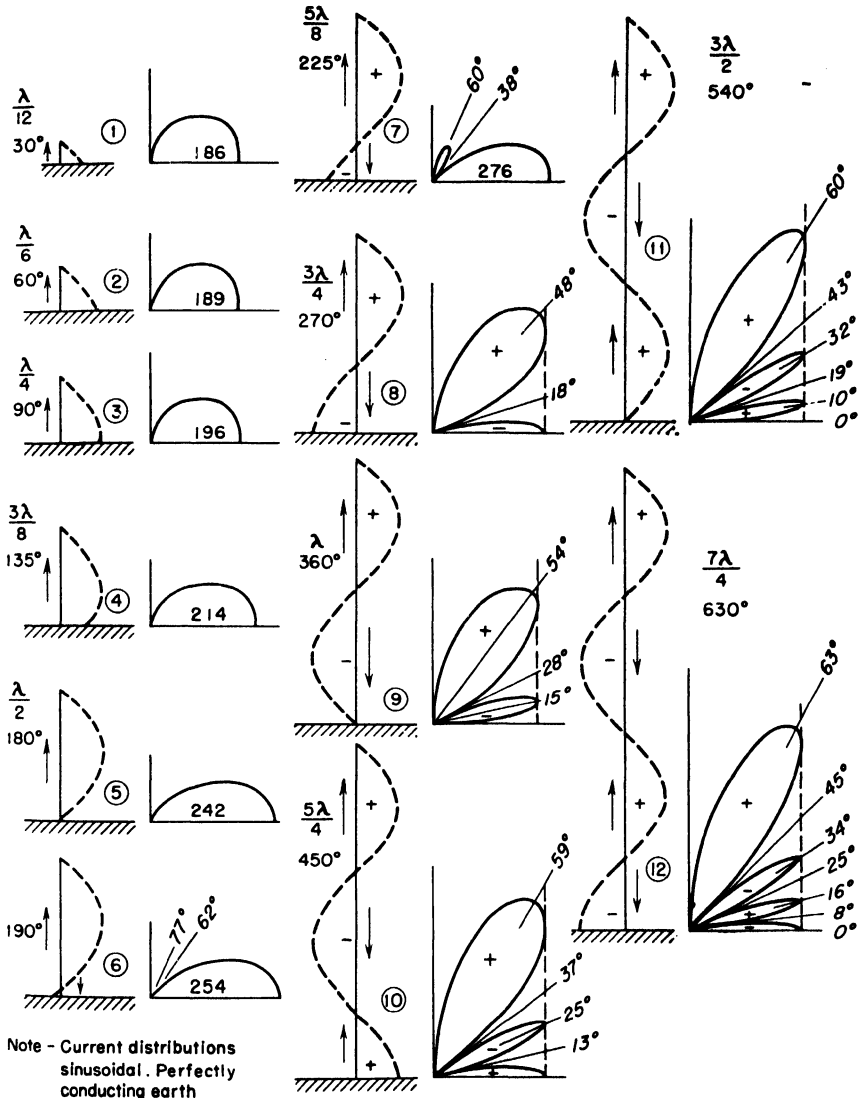



Fig. 12. Vertical field-intensity patterns for vertical antennas. The number shown in the patterns is the effective field in the horizontal plane at 1 mile for 1 kw radiated.

tion, for different vertical angles. The over-all result of this interference in the half space above ground is called the *vertical radiation pattern.* For typical broadcast applications it is given in terms of relative or actual field intensities. Antenna lengths which slightly exceed $\lambda/2$ of sinusoidally distributed current produce a vertical radiation pattern flattened toward the horizon, and this vertical directivity causes a gain in field intensity along the ground. Starting from very short length, the pattern of which is a semicircle through any vertical plane through the axis of the antenna, the

vertical pattern changes very slowly with height until it exceeds $\lambda/4$. Between $\lambda/4$ and $\lambda/2$, the pattern flattens considerably. Greatest field intensity along the ground occurs with a height of $5\lambda/8$ accompanied by a secondary high-angle lobe which is undesirable for its effect on short-distance sky-wave interference with the ground wave at night.[1] Beyond $5\lambda/8$, the ground lobe rapidly contracts, and energy is transferred to the high-angle lobe and thus becomes unsuitable for broadcasting.[2] Thus, from the standpoint of horizontal gain, $5\lambda/8$ represents a maximum condition for natural current distributions. If special means are taken to prevent a reversal of currents below the upper half wavelength of the radiator, further horizontal gain can be

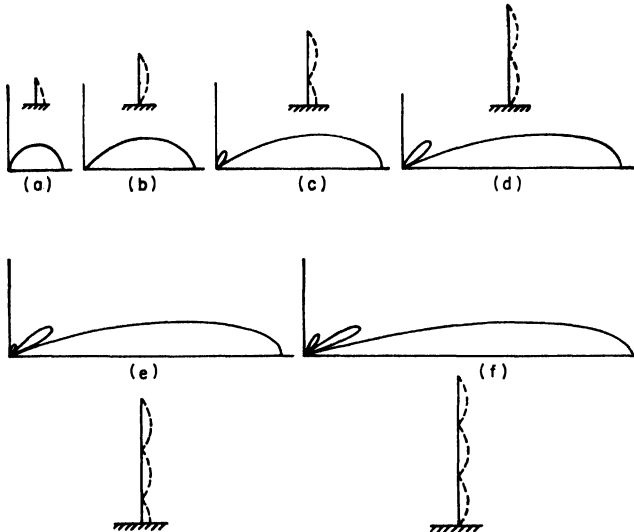



FIG. 13. Polar patterns of the field-strength distribution for vertical antenna over perfect earth when the currents in successive dipole sections are cophased.

obtained with increasing antenna height. This scheme, due to Franklin, is exemplified in Fig. 13.

The dependence of the vertical radiation patterns on current distribution and antenna height is shown in Figs. 12 and 13. In Fig. 12 diagrams 1 and 2 conform to the polar equation

$$f(\theta) = m \sin \theta$$

where m = proportionality constant = effective field at 1 mile for 1 kw radiated
θ = polar angle measured from the axis of the antenna

Diagram 3 has the equation

$$f(\theta) = \frac{m \cos (90^\circ \cos \theta)}{\sin \theta}$$

All diagrams conform to the more general equation derived from a sinusoidal current distribution in a vertical antenna whose lower end is at ground level, and for total electrical length G deg,

$$f(\theta) = \frac{m \cos (G \cos \theta) - \cos G}{\sin \theta (1 - \cos G)}$$

The angles shown in Fig. 12 are the complements of the angle θ.

[1] See reference 20.
[2] See references 70 and 118.

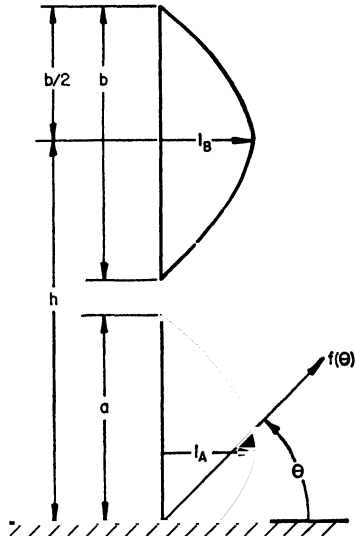
Values of the coefficient m in millivolts per meter over perfectly conducting earth at 1 mile with 1,000 watts radiated are shown in the various diagrams. It is important to note that for diagrams 1 to 3 (a range of heights most frequently encountered in practice) the value of m increases only 5 per cent in going from very small height to a full $\lambda/4$; in going to $3\lambda/8$ it increases only 15 per cent over that for a very short radiator.[1] These facts have an important bearing on antenna economics but are contingent upon simulating virtually perfectly conducting ground by using a suitable ground system.

The FCC requires that all applicants for new, additional, or different broadcast facilities shall specify a radiating system (antenna and ground system), the efficiency of which complies with the requirements of good engineering practice for the class and power of the station. That is, the licensee must request authority to install a new antenna system or to make changes in the existing system which will meet the minimum-height requirements or submit evidence that the present antenna system meets the minimum requirements with respect to field intensities (these requirements are listed in reference 136).

FIG. 14. Notations used in developing the vertical radiation characteristics.

The primary service area of a broadcast station is important from a service and commercial point of view. The ideal antenna for high-powered broadcast stations that provide both ground-wave and sky-wave service should have a vertical radiation pattern such that the sky-wave signal does not interfere with the desirable ground-wave service. The nighttime primary service area would then be practically as large as the daytime area since the fading or distortion wall would not be the limiting factor. It is further desirable that the sky-wave signal strength rise rapidly in order to limit the intense fading area to a narrow band. An antenna that radiates practically no energy at angles above 50 deg from the horizon is shown in Fig. 14.[2] This angle can be varied from 40 to 60 deg with only a small increase in high-angle radiation by changing the current ratio $m = I_B/I_A$. At high angles, the radiation from the upper section, assumed to be 0.5λ long, is 180 deg out of phase with the radiation at these angles from the lower section. By proper choice of excitation currents the fields can be made nearly equal and will cancel.

The optimum antenna dimensions for minimum radiation above 50 deg using a 0.5λ upper-element length are a lower-element length of 0.334λ (120 electrical degrees) and a center height of the upper element of 0.584λ (210 electrical degrees). For these dimensions the value of the current ratio for a null at $\theta = 50$ deg is $m = 0.69$. The effect of the current ratio on the normalized vertical pattern $f(\theta)$ for this antenna is shown in Fig. 15. The theoretical gain of this antenna is about 14.5 per cent compared with a 0.53λ (190 electrical degrees) antenna, or about 41.5 per cent referred to a $\lambda/4$ antenna. The characteristics of the antenna have been verified using antenna models.

The performance of this antenna is compared with that of a 0.53λ antenna in Fig. 16. The sky-wave signal values are for 50 per cent of the time. The center of the fading zone is moved from 127 to 164 miles. The width of the 2:1 sky-wave to ground-wave signal ratio has been reduced from 36 to 24 miles.

[1] See reference 121.
[2] See reference 144.

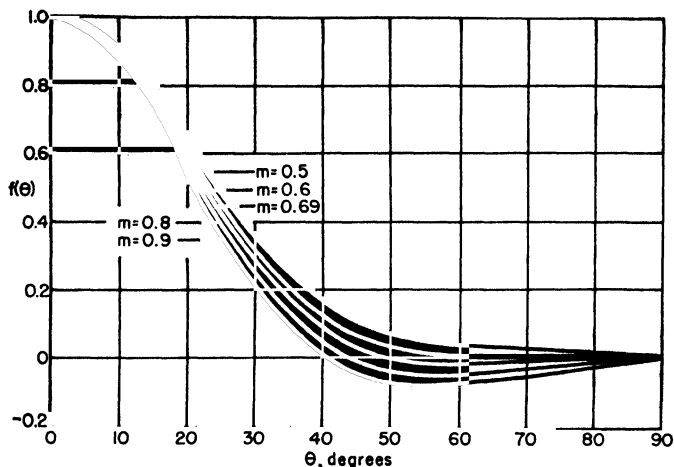



FIG. 15. Vertical-radiation characteristics for various current ratios m, when A is 120 deg, B is 180 deg, and H is 210 deg.

23. Obstruction Lighting. When an antenna system is designed, the regulations regarding the lighting of high structures that constitute obstruction to aviation should be carefully investigated and the necessary details anticipated in the design. Regulations for lighting and painting vary in different countries and with remoteness from airports and established airways. Lighting circuits often must include flashing beacons and spare lights with automatic transfer in case of failure. In some types of antennas the design is seriously affected by the needs of obstruction lighting. Since the antenna system must function correctly with the lighting circuits, all operating measurements must be made accordingly.

When towers are used as radiators, lighting power must either pass through all the coupling and tuning devices or by-pass them through double-conductor chokes of very high r-f impedance or through special toroidal transformers with air spacing between windings, such as the Austin tower lighting transformers.

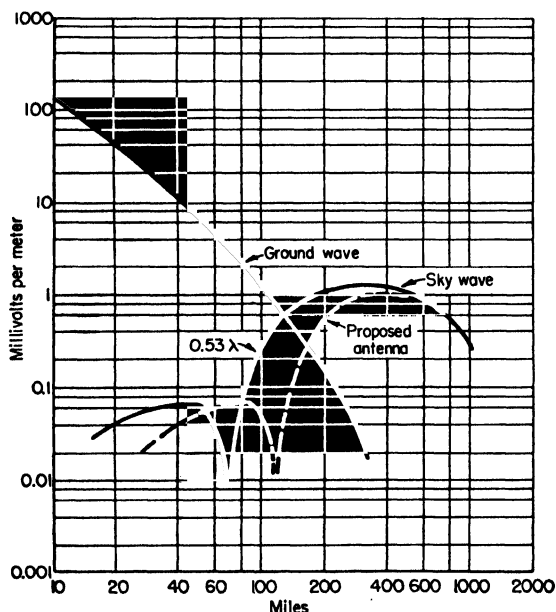



FIG. 16. Comparison of the fading range of a 0.53λ antenna and the proposed antenna. Both antennas produce a field of 1,770 mv per m at 1 mile at 1,000 kc over earth with a conductivity of 10^{-13} emu.

The latter has the characteristic of low capacitance and very high flashover potential between windings. Several versions are available, including one built into a safety core base insulator and immersed in transil oil and capable of withstanding extremely high r-f potentials.

24. Bonding between Sections of Tower Radiators. Special care is required during construction to ensure good electrical contact between tower sections. Good electrical bonding may be required across the joints between vertical sections. To aid in providing good electrical conductivity throughout the length of a tower, galvanizing of the steel is desirable. Towers used as supports do not require galvanizing except in situations where painting is not required for aviation obstruction marking, in which case the spelter provides the necessary protection against corrosion. Where painting is required, the paint will serve as adequate protection when renewed at intervals specified to keep the colors bright. For a tower radiator the paint goes on over the galvanizing.

25. Ground Systems.[1] Efficient antenna systems for operation at these frequencies are usually designed to work with a radial-wire ground system buried just below the ground surface to provide a low-loss return path for antenna base currents. A ground system consisting of 120 buried radial wires (depth of burial sufficient for protection only) $\lambda/2$ long provides a virtually perfect grounding terminal. This length is sufficient to collect earth currents due to the antenna field above ground over the top of the ground wires, and only a negligibly small amount of earth current is collected from beyond the ground wires. Such a system provides a radiation efficiency in typical cases of 95 to 98 per cent in terms of short-distance field intensities realized. A further advantage of this system is with antifading antennas, where the high-angle radiation pattern retains its theoretical form due to the very small reflection losses to waves reflected from the area of the ground system, which functions as a perfectly reflecting surface.

To obtain the maximum efficiency of which an antenna is capable, a good ground system (or counterpoise) must be employed. The FCC sets forth that, at the present development of the art, the ground system for a grounded vertical antenna in the standard broadcast band should consist of buried radial wires at least $\lambda/4$ long. There should be as many radials evenly spaced as practical and in no event less than 90 (120 radials of 0.35 to 0.4λ long equally spaced is considered an excellent ground system, and in case of high base voltage, a suitable base screen should be employed).

In radial ground systems as large as these, the electric flux from the antenna terminates on the ground wires from above, and the current densities under the wires are very low. For shorter wires, where earth currents return to the system from beyond the wires, ground rods at the ends may be justifiable, as with l-f systems. There is no need for circumferential bonding of the radials.

Directive antennas employing two or more radiators require a separate radial ground system for each up to the points where the wires of one system cross those of another, at which points they can be bonded and the projections eliminated.

The antenna input impedance is influenced by the number and length of the radial ground conductors and by the ground conductivity. The effect of wire size is slight. A sensible design criterion for an optimum ground system is attained by a suitable choice of the number and length of ground-wire radials so that they will always carry an appreciable fraction of the total earth current.[2]

A procedure for designing an optimum radial ground system which will assure l-f (below 1,000 kc) antenna efficiencies of 50 per cent or more provided the soil conductivity is accurately known is given below.[3]

Optimum design of the ground system, when the antenna and transmitter design are fixed, is accomplished when the total annual expense of such a system is at a minimum. This includes amortization cost of copper and its installation and the cost of the power

[1] See references 121 and 29.
[2] See reference 166.
[3] See reference 115.

dissipated in the ground. Copper costs are determined by wire diameter and spacing. If K_1 represents the annual cost of the copper wire and installation per meter, then K_1/d is the yearly cost per square meter of surface area, where d is the wire spacing. If C_a is the annual cost of station operation, C_b is the amortization cost of the station exclusive of ground system, and W is the average antenna input power, then $K_2 = (C_a + C_b)/W$ is the annual cost of operation per watt of antenna input power. If w is the power dissipated per square meter of ground surface with a conductor spacing d, then K_2w is the cost of the power dissipated in the ground. The power dissipated per square meter of ground surface is determined by the product of the square of the radial component of the surface current density J^2 set up on the surface of the ground by the antenna fields and the real part of the surface impedance of a square meter of soil (in which the conductors are embedded) in the radial direction. The total annual cost per square meter of ground system is the sum $K_1/d + K_2w$.

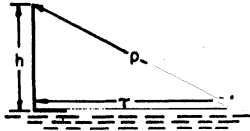



FIG. 17. Notations used in developing the optimum ground system.

The magnitude of the surface current density will decrease with distance from the antenna. At any point on the surface of the ground it is numerically equal to the tangential component of the magnetic field intensity at that point and, therefore, can be measured with a field-strength meter. Or, assuming a simple vertical radiator as shown in Fig. 17, the surface current density can be calculated from

$$\frac{\lambda J}{I_0} = j \frac{1}{2\pi \frac{r}{\lambda} \sin \beta_0 h} (e^{-j\beta_0 \rho} - e^{-j\beta_0 r} \cos \beta_0 h)$$

where I_0 = antenna base current
β_0 = propagation constant = $2\pi/\lambda_0$
λ_0 = free-space wavelength
ρ = distance from top of the antenna to point where J is to be determined
r = distance from base of the antenna to point where J is to be determined
h = height of the antenna

The magnitude and phase of $\lambda J/I_0$ for various antenna heights are shown in Fig. 18 as a function of radial distance from the antenna.

For l-f installations (below 1,000 kc) in nonferromagnetic ground of reasonable conductity (*i.e.*, $\sigma/\omega \gg \epsilon$, where σ = soil conductivity, ϵ = dielectric constant, and $\omega = 2\pi f$), the optimum wire spacing d which gives minimum total annual cost per meter of ground system can be determined from

$$d^3 \left(\log_{10} \frac{d}{2\pi a}\right)^2 = \frac{2.38 \times 10^8 K_1}{J^2 \sqrt{\sigma f^3}\, K_2}$$

where d = optimum wire spacing, m
a = wire diameter, m
J = magnitude of the surface current density, amp per sq m
f = frequency

Since the current density is a function of distance from the antenna, the optimum spacing, and hence total number of radials, will vary with distance. If we let F represent the right-hand side of the above equation, curve I of Fig. 19 gives a plot of the optimum spacing as a function of F for No. 8 AWG wire.

Having determined from Fig. 19 the conductor spacing at several distances r, the number of radials at any distance is given by $N = 2\pi r/d$. Usually the number remains substantially constant and permits selection of a fixed number close to the optimum for all spacings.

The optimum length of the radials can be obtained once the number of radials N is selected and the terminal spacing between conductors d_t is determined. Assuming the same conditions used to determine the above relation for the optimum spacing, the terminal spacing can be determined from

$$d_t^2 \left(\log_{10} \frac{d_t}{2\pi a}\right)^2 = \frac{1.88 \times 10^6}{f\sigma} = F_2$$

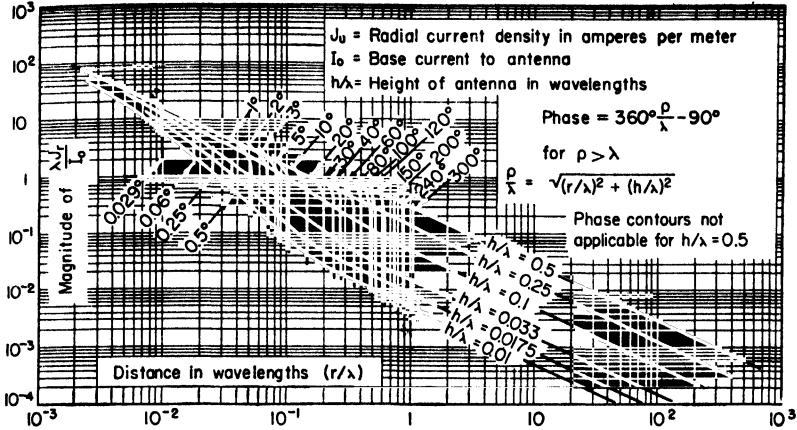



Fig. 18. Phase and density of radial ground current.

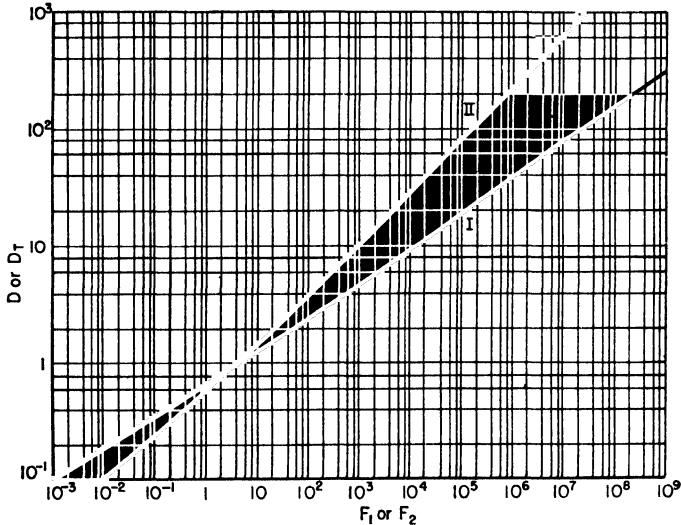



Fig. 19. Optimum and terminal spacing of buried conductors.

Curve II of Fig. 19 is a plot of d_t as a function of F_2 for No. 8 AWG wire. The optimum length of the radial is therefore

$$r_t = \frac{Nd_t}{2\pi}$$

26. Directive Broadcast Antennas. The allocation standards of the North American Regional Broadcasting Agreement pose some difficult design problems for new broadcast stations or those seeking power increases.[1] Most of these require directive antennas of varying complexity. Under the allocation rules, the limiting field intensities at particular azimuth and altitude angles, as determined by geography and wave

[1] See reference 136.

propagation, are calculated, and a directive antenna capable of performing within such limits must be found. A great deal of ingenuity and computational labor is involved in finding the specifications for an acceptable array.[1] To aid in this task, some calculating devices have been developed to shorten the work of cut-and-try approach for three radiators. The RCA Antennalyzer has helped solve pattern problems for two to five radiators in any azimuth, spacing, current amplitude, and phase relationship. With five radiators there are 16 independent variables that must be determined in strictly correct combinations. To do this, the Antennalyzer presents a horizontal pattern in either polar or rectangular coordinates on a cathode-ray oscillograph screen in correct azimuth. The dials are adjusted until a pattern is obtained that meets the limits necessary, and the specifications for the array are read directly from the dials. The pattern can also be made in rms form in a simple operation. Besides its value as an array synthesizer, this instrument is valuable in studying pattern stability problems, since tolerances in any variable can be quickly observed.

Without such an instrument, one proceeds the hard way to find an array giving a pattern meeting prescribed limits of protection and stability with the desired power.

27. Two-element Arrays. A wide range of broadcast directive patterns can be had using two vertical radiators by varying the tower-line azimuth, the spacing, the current ratio, and the phase difference.[2] Such patterns are symmetrical about the vertical plane through the radiators. A two-radiator directive system is the simplest that can be applied to directive broadcasting. Its pattern can be calculated from the equation

$$F(\theta,\phi) = f(\theta)(1 + k_1 e^{j\psi_1})$$

The legend for this equation is given later. For the case of $k = 1.0$ (equal radiator currents), the horizontal field-intensity patterns over a range of values of spacing and phase differences are presented in Fig. 20, which include the effects of mutual impedance. The reference circle in each diagram is therefore the relative field intensity for one radiator with the same power input.

28. Radiation Pattern for Generalized Multielement Array. If the conditions for applying pattern multiplication are satisfied,[3] the radiation pattern for an array of n identical vertical radiators whose ends are normal to a common plane (such as ground), as observed on the surface of an enclosing hemisphere whose radius is very large with respect to the greatest dimension of the array, is given in complex form as follows:

$$F(\theta,\phi) = f(\theta)(1 + k_1 e^{j\psi_1} + k_2 e^{j\psi_2} + \cdots + k_{n-1} e^{j\psi_{n-1}})$$

where $F(\theta,\phi)$ = relative field-intensity pattern in azimuth ϕ and polar angle θ
$f(\theta)$ = relative vertical field-intensity pattern for one radiator
k_m = I_m/I_0 and may be greater or less than unity
$e^{j\psi_m}$ = unit vector at an angle ψ_m with respect to reference vector
ψ_m = phase angle due to space and time differences

and

$$\psi_m = S_m \cos(\phi - \phi_m) \sin\theta + \alpha_m$$

where S_m = spacing in electrical degrees of radiator m with respect to radiator O
= $(360°/\lambda_0) \times$ physical spacing
α_m = phase difference between I_m and I_0; ranges through values of ± 180 deg
ϕ = angle measured clockwise from reference azimuth, usually true north
ϕ_m = azimuth of line through radiator m and radiator O and ranges clockwise through 360 deg

For the horizontal pattern only, $\theta = 90°$ and $\sin\theta = 1.00$. For the vertical pattern through any azimuth angle ϕ, $\cos(\phi - \phi_m)$ = constant.

Each term in this equation represents the phasor contribution of one of the radiators. When all radiators are located on a common line, the radiation pattern is symmetrical with respect to the vertical plane through this line.

[1] See reference 163.
[2] See references 39, 93, 119, and 149.
[3] See reference 9, Chap. 4.

The rms value of a horizontal pattern can be obtained by measuring the area of the polar plot and constructing a circle of the same area.[1] If done in rectangular coordinates, the relative field intensities at each azimuth must be squared, then the area under the squared curve measured. The rectangle that has this same area on the same

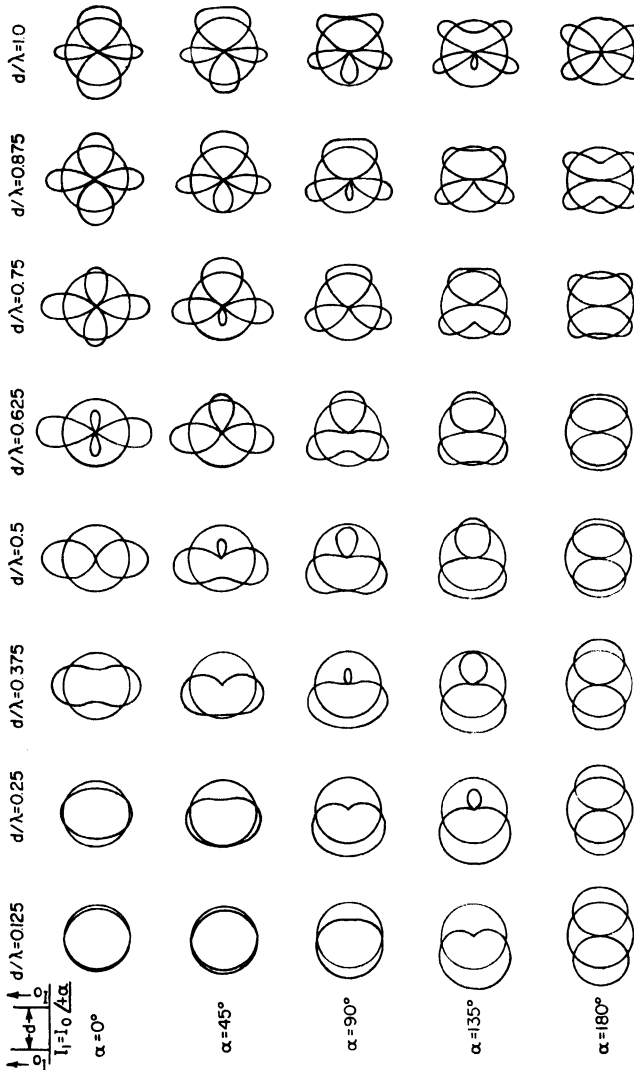



FIG. 20. Horizontal patterns for two radiators fed with equal-magnitude currents taking mutual impedance into account. Circles indicate relative fields from one radiator alone with the same power input.

base then has its altitude squared to give the altitude of the rms field intensity for a single radiator with identical power input. From known performance of single radiators in terms of field intensity, power, and efficiency, the field-intensity scale in absolute values can be determined for the directional pattern.

[1] See reference 164.

29. An Exact Array-design Method.[1] The usual method of array design, which employs pattern multiplication, is an approximation which is sometimes adequate but which never predicts exactly the pattern of a practical array. Consider an array of identical antennas arranged in a uniform lattice. If pattern multiplication is applied, the pattern of the array is the product of an array factor, which depends only on the lattice arrangement, and the primary pattern, which depends only on the type of element used to make up the array. This procedure usually produces very good results when applied to arrays of dipole antennas. The error in this procedure is in the assumption that the radiation pattern of each element (when radiating in the presence of the remaining elements) is the same as the radiation pattern of an isolated element (when radiating in the absence of all other elements). This is never true in practice, although it may be very nearly true in certain cases. The distortion of the primary pattern of any element is caused by the scattering or diffraction of radiation by the other elements in the array.

An exact design method can be stated as follows. Let P_n represent the radiation pattern (amplitude and phase) obtained when unit current is injected into the terminals of the nth element and the terminals of all other elements are open-circuited. Let $I_0, I_1, I_2, \ldots$ represent the amplitude and phase of the currents flowing into the terminals of the elements when all the elements are energized. Then it follows from the principle of superposition that the pattern (amplitude and phase) of the array is given by

$$P = I_0P_0 + I_1P_1 + I_2P_2 + \cdots$$

In practice one would use pattern multiplication to obtain a first approximation to the array structure, build it, measure the amplitude and phase patterns P_n of each element in the presence of the entire array structure, and determine the distribution of currents $I_0, I_1, I_2, \ldots$ (by pattern superposition) which gives the best approximation to the desired array pattern. Usually this shows that the array could be improved by a small modification of the structure. When the structure has been fixed, the feed system can be designed according to the exact procedure and then the resulting array pattern will be exactly as predicted.

The exact method can be formulated equally well in terms of the patterns Q_n obtained when unit voltage is applied to the terminals of the nth element with all other inputs short-circuited.

If there is mutual impedance between the input terminals, the patterns Q_n will be essentially different from the patterns P_n; both P_n and Q_n will be affected by the location of the terminals. Thus where mutual impedance between the inputs is significant, one would try to select the position of the input terminals of the elements to make the functions P_n or Q_n as simple as possible. If the mutual impedance between terminals is insignificant, the patterns P_n and Q_n are the same, apart from a constant, and are independent of the loads connected to the unexcited inputs. However, these patterns in general will differ from the pattern of a single element with all other elements removed.

30. Feeder Design of Broadcast Arrays. The physical realization of a calculated directive antenna involves great skill in design and adjustment and great heed to accuracy of details. Through radiation couplings, any change in spacing, azimuth line for the radiators, current ratio, and phase difference of any radiator current reacts upon every part of the entire radiator and feeder system. Even with simple two-radiator systems, cut-and-try methods of adjustment are virtually hopeless. It is necessary therefore to predict accurately the conditions which must exist throughout the feeder system when working as desired and, by construction and careful measurement, to provide those conditions precisely. This naturally becomes more difficult as the number of radiators increases. In symmetrical systems there is a helpful simplification of work by dealing with radiators in pairs. The use of identical radiators of accurately known self-impedance and current distribution is particularly important as a starting point. For close-spaced radiators of the antifading type, the current distribution is modified by mutual coupling, which in turn modifies the expected feed-point impedances. Care must be exercised that guys on radiators do not measurably

[1] See reference 51.

deform the basic circularity of pattern for each radiator. Ground systems must present identical resistances in each radiator.

In designing a directive antenna system, the feed-point impedance of each radiator must be calculated for conditions which will produce the desired radiation pattern.

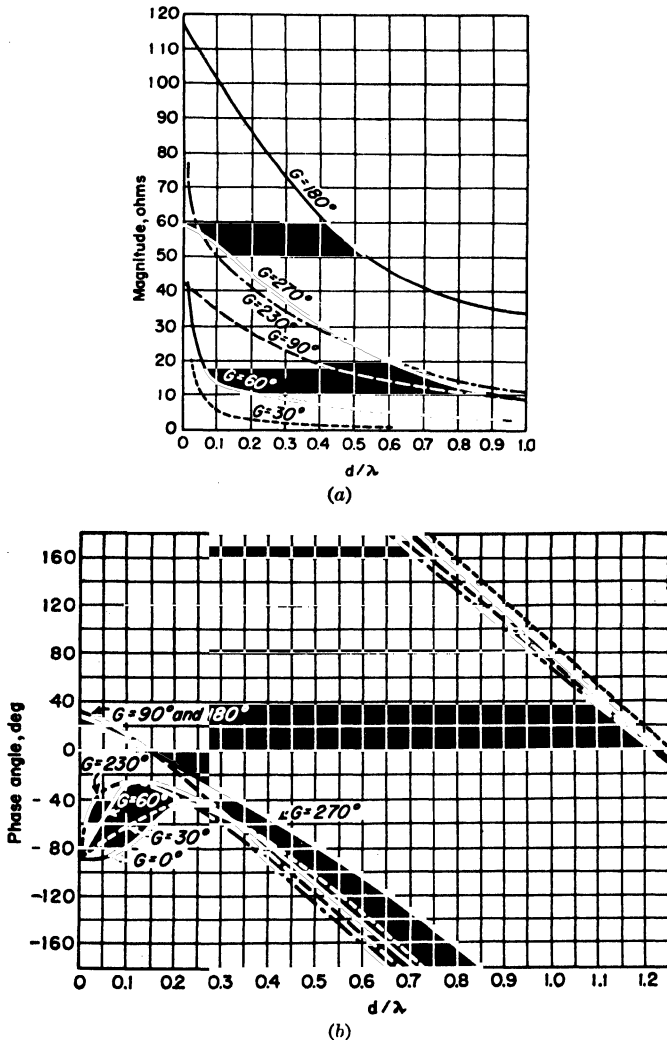



FIG. 21. (*a*) Mutual impedance between antennas. (*b*) Phase effect of coupled antennas.

This is done as follows. Let $Z_{00} = Z_{11} = \cdots Z_{mm}$ represent the self-impedance of an element of the array obtained from measurement of one prototype radiator at and about the working frequency, including base insulators, tower lighting devices, phase monitoring circuits, and any other attachments actually employed in final working status. However, identity of radiator impedances is not always assured by identical construction.

The transforming effect of all attachments to the radiator should be known as accurately as possible to be able to correct the mutual-impedance effects as transferred to the feed points, since the available mutual-impedance data are based on ideal conditions without such transforming effects present.

Measured values of Z_{00} (reference radiator) are then converted to their value *inside* the transforming devices. This then provides the idealized values of self-impedance.

Mutual impedances between all combinations of the radiators are then calculated or read from curves.[1]

The magnitude and phase of the mutual impedance between grounded vertical antennas of equal electrical height G are given in Fig. 21a and b as a function of antenna spacing d in wavelengths. The resistive and reactive components of the mutual impedance between vertical antennas of unequal heights located above a perfectly conducting ground assuming sinusoidal current distributions are shown in Figs. 22 to 25 as a function of antenna separation d for typical combinations of antenna heights.[2]

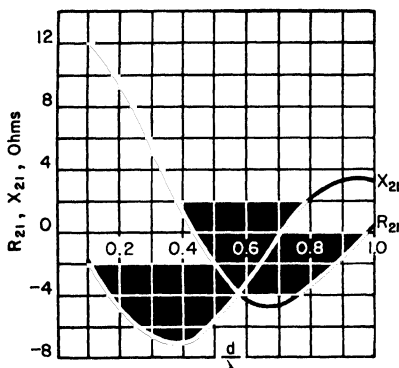



FIG. 22. Resistive and reactive components of mutual impedance between antennas of heights of 40 and 90 deg.

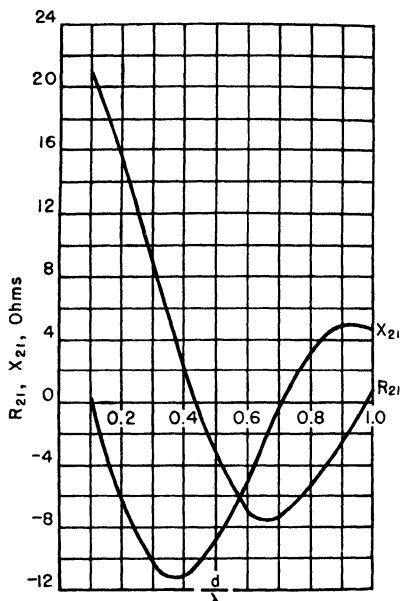



FIG. 23. Resistive and reactive components of mutual impedance between antennas of heights of 60 and 90 deg.

Simultaneous equations are then set up including the complex current relationships required in the various radiators, in the idealized location beyond the transforming effects:

$$
\begin{aligned}
E_0 &= I_0Z_{00} + I_1Z_{01} + I_2Z_{02} + \cdots + I_mZ_{0m} \\
E_1 &= I_0Z_{10} + I_1Z_{11} + I_2Z_{12} + \cdots + I_mZ_{1m} \\
E_2 &= I_0Z_{20} + I_1Z_{21} + I_2Z_{22} + \cdots + I_mZ_{2m} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
E_m &= I_0Z_{m0} + I_1Z_{m1} + I_2Z_{m2} + \cdots + I_mZ_{mm}
\end{aligned}
$$

The vector ratio of E_0/I_0, E_1/I_1, E_2/I_2, etc., for each radiator then gives the complex impedance (idealized) for each radiator. These impedances are then transformed back to the actual feed-point impedances (taking into account any phase shifts introduced by the transformation) for the several radiators.

Then the power input to each radiator must be calculated for the currents used in the previous calculation (based at this stage on current ratios) and the division of

[1] See references 32, 58, 59, and 119.

[2] See reference 25.

power among the several radiators determined. With power distribution known and also the feed-point impedances, the feed-system and the coupling-phasing networks can be calculated for each radiator and for the power-dividing networks. The designer has a wide latitude of choice in the arrangement of feeders, the type and characteristic impedance to use, and in the location of the power-dividing networks.

Owing to the fact that mutual-impedance vectors can lie in any of the four quadrants,[1] it is possible to find that, in some circumstances, the input impedance to a radiator has a negative resistance. This of course means that instead of feeding power into this radiator, a certain amount of power is delivered from the radiator into the feeder system. This is negative power from the standpoint of determining power division, and the feeder line must be terminated at the opposite end. Wherever possible to do so, it is desirable to avoid negative resistance inputs to radiators in the interests of system simplicity. However, in complicated arrays this may be unavoidable.

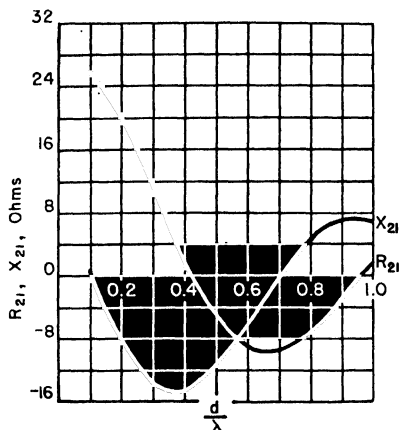



FIG. 24. Resistive and reactive components of mutual impedance between antennas of heights of 75 and 90 deg.

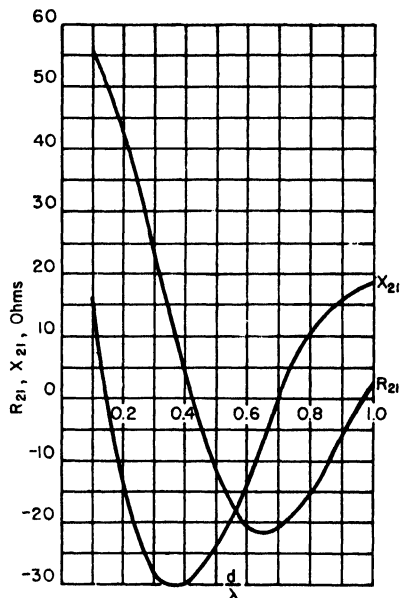



FIG. 25. Resistive and reactive components of mutual impedance between antennas of heights of 120 and 90 deg.

It is customary to terminate all feeders so as to suppress standing waves. Then the phase lag in a feeder is equal to its electrical length, which in turn depends upon the velocity of propagation in the lines used. If open-wire lines are used, the velocity is the same as that in free space. Enclosed lines always have velocity reduced in some degree, and this must be accurately known in order to use the electrical length of the line for the phasing calculations.

It is not necessary to suppress standing waves in the feeders unless they are so high as to cause appreciable power loss or high voltages. When not terminated, the line acts as a transformer, the input impedance is generally complex, and the phase lag for the line current is not the same as the electrical length of the line (see Fig. 26).

In a directive antenna it is the current that must have proper phase and amplitude. Thus, in the design of the feeder and coupling system, the current phases are to be controlled. Starting with the transmitter, the total phase difference to each radiator, the power-dividing networks, feeders, and coupling networks must be such as to bring to each radiator its correct relative phase and amplitude. Most of the networks

[1] See references 119, 58, and 59.

will have to make a specified impedance match with a specified phase shift, and networks must be individually synthesized on that basis.

When the system has been fully calculated and constructed ready for adjustment, the individual elements of each network are adjusted to calculated values at the working frequency with suitable measuring instruments.[1] Inductance adjustments must include external leads. After this operation, the network can be further verified by connecting an impedance across its output terminals equal to the anticipated working load and having its input impedance measured, trimming adjustments until correct input impedance is noted. Such trimming should involve only very small changes or

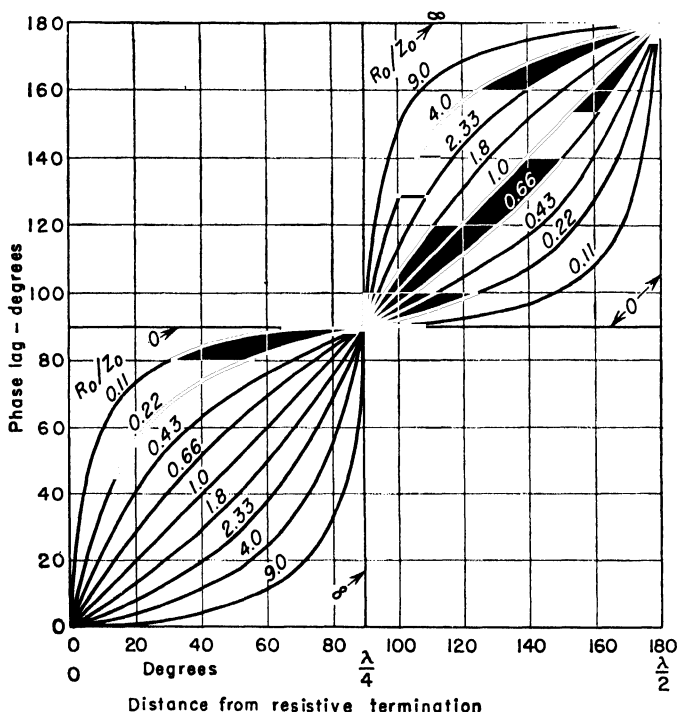



FIG. 26. Phase lag vs. line length for dissipationless line with resistive termination for various ratios of load resistance to characteristic impedance R_0/Z_0.

otherwise the phase-shift characteristics of the network may be incorrect even though the impedance transformation is correct.

When connected completely, the entire system can be energized with power, and the various currents in the system checked for amplitudes as a first test. If correctly calculated and measured, very nearly correct relative values should result immediately. Next, phases are checked by a phasemeter, or by the ultimate test, the measurement of the pattern. A first measurement of field intensity is made to determine that the nulls and minimums occur at correct azimuth angles. Careful circuit trimming ensues until the ammeters (previously checked for correct calibration) show correct relative values and the pattern nulls occur at correct angles.

Final test is to measure carefully the entire horizontal pattern at points of small enough azimuthal difference to provide full data in relation to the complexity of the

[1] See reference 119.

pattern. Measurements must be made on each angle to permit calculation of the unattenuated field intensity at some constant distance. For large arrays this may be the 2- or 3-mile circle, so that the pattern is fully formed. If the pattern is found to have the correct shape and orientation, power adjustments are then made for rated input, upon which basis the proof of performance is made.

31. Stability and Distortion in Directive Antenna Systems. Stability includes the effects of changes of the design parameters as well as the frequency. Many theoretical designs become impractical when viewed from the standpoint of stability. All system connections must be thoroughly secured and network-connecting leads held rigidly in their tested positions. Effects of corrosion, weather, deterioration, heating, etc., must be carefully anticipated to assure stability of adjustment for long periods. Networks should be placed in tamperproof locked boxes. With such precautions, directive systems maintain excellent long-term stability.

The impedance of a broadcast antenna, and particularly the common-point impedance of an array, often varies so much over the transmission band as to impair the frequency response, amplitude linearity, and modulation capability of the transmitter.[1] Recognition of this difficulty has led to the formulation of an EIA specification for the "normal load" into which a transmitter should operate and meet its performance requirements. This is a load whose resistance does not depart more than 5 per cent from its carrier-frequency value at ± 5 kc or 10 per cent at ± 10 kc, and whose reactance, which is zero at the carrier frequency, does not exceed 18 per cent of the carrier-frequency resistance at ± 5 kc or 35 per cent at ± 10 kc.

The results of measuring the side-band power, percentage modulation, and distortion of the signal delivered to a sharply tuned antenna system are considerably different at different points in the coupling system or along the transmission line. For example, if the variation in antenna impedance is equivalent to a series-resonant circuit at the carrier frequency, then under certain conditions of audio-input level to the transmitter (at a single audio frequency), the *voltage* at the antenna terminals (and any $\lambda/2$ back from the terminals) will appear to be fully modulated; that is, the magnitude of the two side-frequency voltages will be one-half the carrier voltage. But, because the impedance rises on either side of the carrier frequency, the side-frequency currents will be less than one-half the carrier current and an inspection of the current envelope would show less than 100 per cent modulation. The inverse is true at points an odd $\lambda/4$ from the termination. At points an odd $\lambda/8$ from the termination the impedance-vs.-frequency variation is dissymmetrical and both the voltage and current-modulation envelopes would appear distorted, and measurements at this point would indicate a high percentage of distortion in spite of the fact that no extraneous side frequencies are being radiated.

To determine actual delivered side-band power by measuring the current envelope, it is necessary to make the inspection and measurement at a point where the *series* resistance is independent of frequency over the band transmitted since it is the current squared times the resistance that determines power. These points occur every $\lambda/2$ from the termination. If the voltage envelope is to be monitored the point of measurement should be at a point where the *parallel* resistance is constant over the frequency band since the power in this case is the voltage squared divided by the parallel resistance. These points occur every odd $\lambda/4$ from the termination. In the first case, the maximum permissible modulation of the wave being analyzed has to be set by inspection of the voltage envelope at the odd $\lambda/4$ points, and in the second case, the modulation has to be set by inspecting the current envelope at $\lambda/2$ points. To permit plotting conventional curves of distortion vs. modulation frequency for particular percentages of modulation, it is necessary to express the percentage modulation in terms of the quantity (current and voltage) which can be allowed to attain full modulation at the monitoring point, even though the true side-band power, with sharply tuned loads, does not correspond to full modulation. There is justification for this in the fact that, when the maximum permissible modulation is reached, the power-amplifier tubes are

[1] See reference 129.

being required to deliver either full peak current or full peak voltage to the load, even if not both.

The point of connection of the power-amplifier tubes is of profound importance in the performance of the amplifier regardless of the circuit or modulation method used. Since one can ask only that a transmitter deliver a voltage or current wave free of phase modulation and having its envelope identical in shape to the audio-input wave, then only a point of impedance symmetry such as the $\lambda/2$ and odd $\lambda/4$ points is appropriate for making connection to the tubes. If envelope feedback is used only the odd $\lambda/4$ points are satisfactory.

The design of a variable-phase shifter or *line stretcher* with a total range of about 180 deg for use in broadcast transmitters to obtain the required impedance-vs.-frequency characteristics is shown in Fig. 27. In this figure $\omega L_1 = \omega L_2 = R$ = terminating resistance, $\omega C_1 = \omega C_3 = [1 - \cot(\Phi/2)]/R$, and $\omega C_2 = 2[1 - (\sin \Phi)/2]/R$, where Φ is the desired phase shift. In some cases it can be inserted between the r-f plate terminal of the power amplifier and the transmission-line matching section. This network has good harmonic-suppression characteristics.

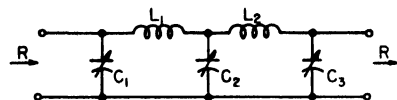



FIG. 27. Phase shifter for use in broadcast systems.

In addition to the signal distortion introduced by the frequency selectivity of the antenna input impedance, a second type of distortion can result from the directional characteristics of an antenna array.[1] For example, the directional pattern of an array is a function of the transmitted frequency, and can be different at the side-band frequencies from that obtained at the carrier frequency. These radiation-pattern differences cause changes in the relative magnitudes and phases of the signal components and hence cause distortion of the modulation envelope. At low audio-modulating frequencies the directional properties of the pattern for the carrier-frequency and the side-band-frequency components are essentially identical. However, at high audio-modulating frequencies and relatively low carrier frequencies the pattern band width for a particular directional characteristic may be sufficiently narrow as to result in severe changes in the relative amplitudes and phases of the signal components. In the null direction, for example, the carrier signal may be completely canceled, but a side-band component may reach reinforcement. In the directions of maximum carrier signal the converse conditions are likely to occur, resulting in reduced high-frequency side-band power.

Arrays of many elements, high-Q tuning networks, negative power elements, and deep nulls are more likely to exhibit severe antenna distortion than simpler, lower-Q antenna systems. Deep nulls should be avoided if service is to be rendered in the null directions.

32. Half-wavelength Bent-wire Antenna.[2] There are other services besides a-m broadcasting that use the m-f band. Some of these services can use broadcast-antenna designs, but for others, wire antennas are more practical. A simple yet effective method of obtaining a high radiation resistance at the feed point of a grounded vertical antenna is shown in Fig. 28. The horizontal section should be $\lambda/4$ long so that a current maximum will appear at the top of the vertical section. Since the most useful radiation takes place from the vertical section at these frequencies, it should be as long as possible, preferably $\lambda/4$. In any case, the horizontal section should be $\lambda/4$ long.

FIG. 28. Half-wave bent-wire antenna.

A parallel-resonant coupling circuit should be used if the vertical section is greater than about $\lambda/8$. For vertical lengths less than $\lambda/8$ a series-resonant coupling circuit should be used.

[1] See reference 156.
[2] See reference 3.

Antennas of this type offer an opportunity for using a simple feeder system that permits installing the antenna some distance from the transmitter. If the vertical section is $\lambda/8$, for example, the total radiating length is $3\lambda/8$. An additional $\lambda/8$ may be added to the antenna to make the total length $\lambda/2$. This additional length is used as the feeder and hence should run parallel and close to the ground to reduce the radiation from it. A parallel-resonant coupling circuit should be used between the input end of the feeder and ground.

In all cases the distance between the coupling circuit and the effective ground should be only a few feet. These leads are part of the effective length of the antenna, and must be added to the antenna length in determining the constants required in the coupling circuit.

Since ground-wave propagation or low-angle radiation for sky-wave propagation is of prime importance at these frequencies the vertical portion of any top-loaded vertical antenna system should be as long as possible. The horizontally polarized radiation from the horizontal section of the antenna occurs at high angles and contributes essentially nothing to the ground wave, and in the daytime the sky wave is absorbed by the ionosphere. At nighttime, however, this horizontally polarized sky wave is reflected with little attenuation. Horizontally polarized radiation from the bent-wire antenna, for example, can be minimized by folding the horizontal section back on itself.

33. Antennas on Ships. Ships of all sizes, with their masts, rigging, derricks, etc., severely circumscribe the antenna designs which can be used. As time goes on, more and more antennas are needed for new radio services on ships. There is usually the "main" antenna for l-f transmission and reception, one or more h-f antennas, a sense antenna for the direction finder, an antenna for the ship's broadcast receiver, and occasionally others for special uses. Their disposition depends largely upon the architecture of the ship and the location of the various facilities and operating rooms. They are located wherever space can be found.

The main antenna is usually a single-wire inverted-L or T antenna suspended between the main masts. Being electrically short at the lowest working frequencies, there is a rather high potential on the system when used with medium- and even low-power transmitters. Therefore, the insulation must be electrically and mechanically robust throughout its length and through the deck bushing to the transmitter. This antenna is often used also for h-f service, in which case the system is electrically very long and provides a many-lobed pattern with vertical and horizontal polarizations. The random radiation patterns obtained provide the necessary functions for communication. The same is essentially true for the single-wire antennas installed for h-f duty only.

The sense antenna for use in conjunction with the direction-finder loop antenna is as nearly vertical as can be and terminates at the d-f location.

As a consequence of typical shipboard circumstances, there is little or no electrical designing possible in the sense in which it is applied to types of land services. There are important mechanical design considerations to be met to withstand the severe weather encountered at sea.

34. Antennas on Aircraft. Only a few elementary remarks will be made on this complicated subject, and these apply to the simplest applications on private and transport aircraft. The characteristics of aircraft antennas are largely empirical, depending upon the frequency, the size and shape of the aircraft, and the available locations. For slow flying speeds, fixed antennas are the rule and so disposed as to offer a minimum drag during flight. Thus antennas are generally oriented parallel to the slip stream over the fuselage (for l-f and h-f services) from a short mast over the cockpit to the tail structure, and under the fuselage for the 75-Mc marker antennas. Vibrational stress is severe, and precautions must be observed in choice of material and method of rigging to sustain typical conditions. At the frequencies where the antenna has high reactance and low resistance, a few watts of antenna power produce very high potentials, which, with high altitude, easily give rise to corona and flashover. Corona formation in high-velocity air occurs at higher potentials than in still air. Care is

required to avoid sharp points which accentuate discharge, and the conductors should be as large as possible. Insulated antenna wire raises the corona potential and also reduces precipitation static.

Direction-finder loops are sometimes fully exposed in low-speed ships, enclosed in streamlined envelopes for medium-speed ships, or located inside the nose of the ship behind a dielectric envelope. To reduce the size of exterior loops and thus decrease drag for exterior loops, powdered-iron cores can be used. When two loops are used with automatic direction finders, exterior mounting is usually necessary.

For instrument landing, both localizer and glide-path antennas are usually located somewhere near the cockpit or the radio room. For v-h-f operation these are presently small horizontal single-turn fixed loops with balanced outputs to the equipment. Vertically polarized v-h-f communication equipment often uses a short whip antenna. U antennas can also be applied to v-h-f reception of horizontal polarization. Flush-mounted antennas are required for high-speed and supersonic aircraft and missiles. Practical flush-mounted u-h-f and s-h-f antennas have been developed for this service and are described later in this chapter.

Radio altimeters employ a pair of electric-dipole antennas, one located under each wing with the dipole parallel to the slip stream. Careful streamlining of the antenna is required for this application, and a spacing insulator is used at the center for mechanical rigidity.

Trailing wire antennas are impractical for high-speed aircraft unless they are very short. A trailing wire v-h-f antenna, with coaxial transmission line and a coaxial dipole antenna at its end for the antenna and feeder system from the rear of the ship, may be used in some circumstances. For private aircraft, a prevailing type of h-f trailing antenna passes from the antenna reel through a guide at the top of the tail structure, and is held out in the slip stream by a small wind sock at its end. When reeled in, this antenna is secure and obscure.

H-F ANTENNAS (3 TO 30 MC)

35. Straight Horizontal Half-wave Dipole. A straight conductor parallel to earth, of the order of one-half wavelength long, suitably excited, is one of the most elementary and useful h-f radiating systems. When exactly $\lambda/2$ long, and in free space, a thin wire antenna has a center-point impedance of approximately $73 + j30$ ohms. When placed parallel to and within about 5 wavelengths of ground, mutual impedance with its image causes the impedance to vary appreciably with its height. Figure 29 shows how the radiation resistance varies with height for both horizontal and vertical half-wave dipoles.[1]

The electric field around a $\lambda/2$ dipole resembles that of the magnetic field of force around a short bar magnet. When in free space, the pattern factor at constant distance follows the relation

$$F(\theta) = \frac{\cos(90° \cos\theta)}{\sin\theta} \qquad (\theta \text{ measured from wire})$$

from which it can be seen that there is zero field intensity in the direction of the dipole and maximum everywhere at right angles to it. It is therefore a directive radiating system. In practice and when used over ground, some endwise radiation occurs because of reflections from ground and from the ionosphere, and in this direction the field is vertically polarized. Normal to the wire the field is horizontally polarized. In intermediate directions there are components of both.

In h-f applications the most important characteristic of such an antenna is its vertical-plane pattern normal to the wire. Over perfectly conducting earth, this has the equation (origin at dipole above ground)

$$F(\phi) = \cos(H \sin\phi \pm 90°)$$

[1] See reference 32.

When the complex reflection coefficient of the ground is taken into account, the equation becomes

$$F(\phi) = 1 + Ke^{-i2H \sin \phi}$$

where $F(\phi)$ = relative field intensity as a function of elevation angle ϕ
H = electrical height of antenna above ground = $2\pi h/\lambda_0$
K = complex reflection coefficient for horizontal polarization and is a function of soil conductivity σ and inductivity ϵ, the angle of incidence ϕ and the frequency f, derived from the equation

$$K = \frac{\sin \phi - (\epsilon^{-i2\sigma/f} - 1 + \sin^2 \phi)^{1/2}}{\sin \phi + (\epsilon^{-i2\sigma/f} - 1 + \sin^2 \phi)^{1/2}}$$

The phase angle is always 180 deg in this case.

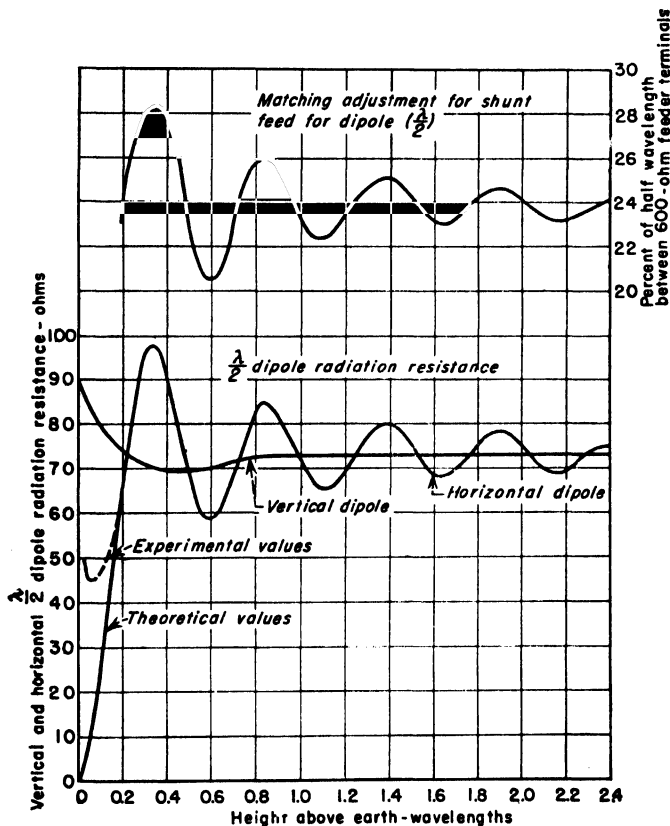



FIG. 29. Radiation resistance vs. height.

For perfectly conducting ground (hypothetical), $K = 1$. Figure 30 shows a series of these patterns for various heights h up to $2\frac{1}{2}\lambda$. The relative polarity of the field in each lobe is indicated. Figure 31 charts the angles of the maximums and the nulls for these patterns. These angles are important for propagation purposes. Figure 32 exhibits the optimum radiation angles for propagation via the various ionosphere heights as functions of distance for one-hop transmission.

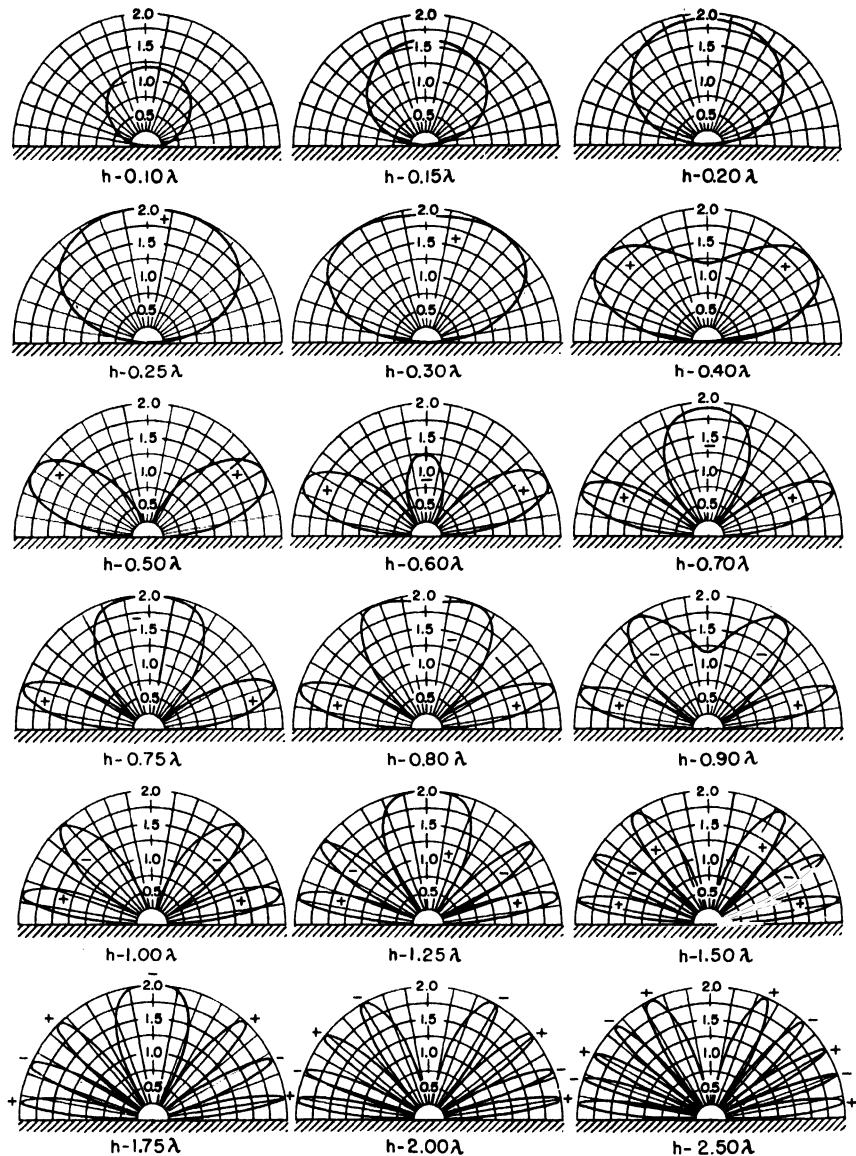



FIG. 30. Vertical polar diagrams for horizontal dipoles.

Both transmitting and receiving antennas of this type, used over fixed circuits, should be at the same height and perpendicular to the direction of transmission.

36. Feed Systems. The straight half-wave dipole can be center-fed in series from a balanced feeder, or shunt feed (also called delta- and Y-feed).[1] The shunt-feed system, when correctly performed, makes a satisfactory impedance match directly

[1] See reference 32.

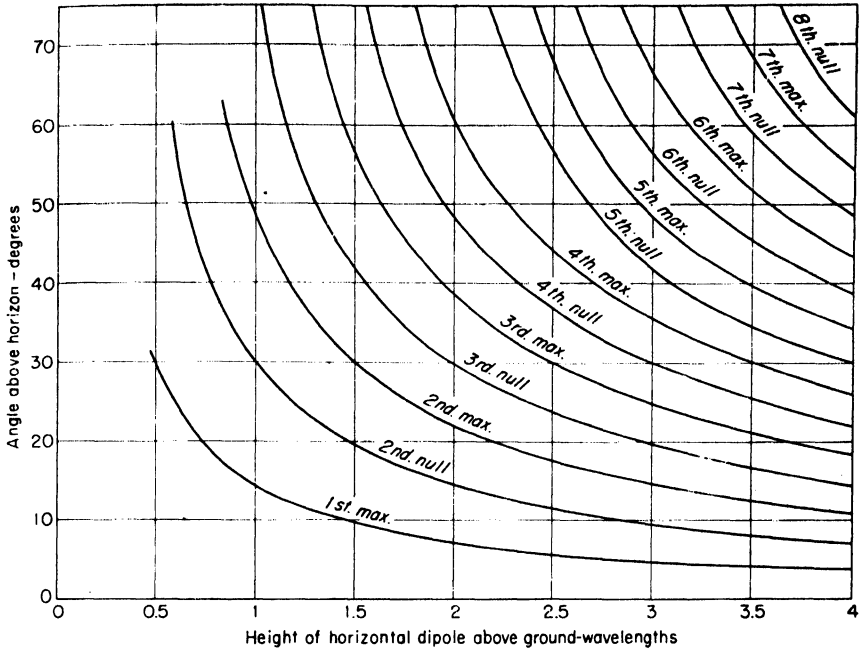



Fig. 31. Angles of maximums and minimums, horizontal dipole.

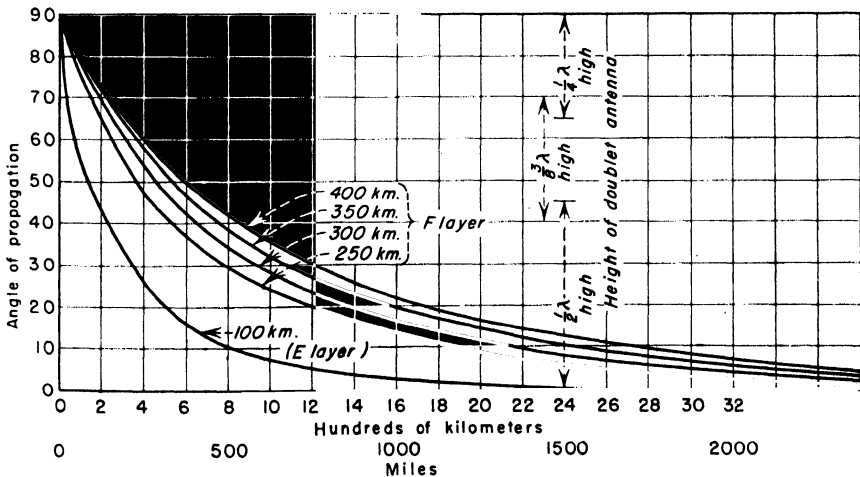



Fig. 32. Optimum radiation angle for propagation for one-hop transmission.

(but never perfectly, unless series capacitors are included to correct for the inductive effect of the enclosed portion of the antenna) for low standing-wave ratio on the feeder. Care must be taken to maintain exact symmetry of connection between line and antenna. Adjustments for this are shown in the upper part of Fig. 29.

Feeders should always run normal to the antenna wire for as far as possible from

the antenna, and skew relations avoided wherever possible. Skew relations cause radiation couplings between feeder and antenna which upset line balance and compromise pattern and efficiency. This is quite general and should be observed in all antenna designing.

Instead of spreading the feeder to make an impedance match, the dipole can be shortened and a small inductance inserted between the junctions with the feeder. The latter is run at fixed separation directly to the antenna. A compromise match can be made in this way, in which case the antenna is then known as a *shortened dipole*. Its advantage over the simpler arrangement preceding is nil, and the operation of adjusting is more difficult.

Another feed method, technically inferior but often used for its simplicity, is off-center feed, with a one-wire feeder, working against ground.[1] A position can be found on the antenna where standing waves on the feeder are quite small for one frequency.

Transformer coupling between feeder and antenna can also be used but there is no good reason to do so.

Both the input impedance and the current distribution (and hence radiation pattern) of dipole antennas (from $\lambda/2$ up to 3 to 4λ long) depend upon the location of the feed point as well as the over-all length of the antenna.

The current distribution on linear-dipole antennas is the sum of attenuated traveling waves influenced by mutual coupling between its arms and its image if located above ground. For thin short dipoles, the distribution is essentially sinusoidal, while for thick and/or long dipoles the distribution is essentially a traveling-wave distribution. This is evident from the radiation patterns.[2] The effect of feed location and antenna length and thickness on the input impedance and radiation pattern can be found in the literature.[3]

It is possible to feed an antenna at a current loop not necessarily at the center. The impedance will be higher at these loops. The so-called "Windom" antenna is a half-wave dipole fed off center with 300-ohm line and as such works well at the higher harmonics of the wire. The SWR will be higher than when fed with 72-ohm line at the center, but if the leads of the transmission line are closely spaced, as with 300-ohm Twinlead, the radiation from the line will be low.

Another type of all-wave antenna is a shortened dipole with tuned *LC* traps inserted each side of the center at such a distance that the center portion is resonant at the second harmonic. The traps are tuned to this harmonic frequency. Such a system will operate with low SWR on all harmonic frequencies of the fundamental.

37. Radiated Field Intensities of Dipole. In free space, maximum field intensity at 1 mile (d = 1,610 m) with 1,000 watts radiated from a $\lambda/2$ dipole is 138 mv per m. Owing to ground reflections, pattern maxima vary from one to two times this value depending upon height and ground reflectivity.

38. Selectivity and Band Width. The selectivity of a $\lambda/2$ dipole is maximum for thin wires where it can be considered to have a band width of the order of ± 1 per cent of optimum frequency. Where larger band widths are to be transmitted quite uniformly, the ratio of diameter to length of one-wire or cage antenna should be the largest practicable. If propagation considerations permit, the heights should be made such as to give maximum radiation resistances, as shown in Fig. 29. Where band widths in excess of ± 5 per cent in frequency are required, other types of construction should be used such as folded dipoles.

39. Antenna Potentials. Where high power is to be transmitted, or at high altitudes, antenna insulation and conductor designs require heed to details. For h-f use, only radial potential gradients need be considered. At high altitudes, pluming may occur with consequent damage to the system. Fortunately in practice, high power is generally used with directive antennas, and the power is divided among several

[1] See reference 200.
[2] See references 9 (Chap. 9) and 199.
[3] See references 196 and 226.

dipole sections, thus tending to minimize this problem. A thin-wire dipole gives an end potential of about 3,900 volts rms for 1,000 watts antenna input for a height of 0.25λ. It will be higher for smaller heights, and falls to a minimum of about 1,700 volts as height increases to 0.75λ; beyond this point it settles down to the free-space value of about 3,000 volts. Potentials vary as the square root of the power ratio and as the inverse square root of the capacitance per unit length. For a potential of 3,900 volts on a wire 0.101 in. in diameter (No. 10 B&S), the radial gradient is of the order of 31 kv per cm. As a rough approximation for a cage, the gradient for one wire is divided by the number of wires in the cage.

40. Half-wave Dipole for Receiving. In receiving, one wishes to know the receiver input voltage when the receiving antenna is immersed in a plane wave field of E volts per m when both the antenna and the receiver are correctly matched to a feeder of known characteristic impedance Z_0. This can be approximated in the following manner:

The area over which a $\lambda/2$ dipole collects energy is 0.1305 square wavelength.[1] The power density within this area from the field is $P = E^2/377$. If the antenna is in free space and has no losses, the power intercepted would be, for optimum antenna orientation,

$$W = \frac{0.1305\lambda^2 E^2}{377} \qquad \text{watts}$$

so that the receiver input voltage would be

$$V = \sqrt{\frac{W}{Z_0}}$$

Over ordinary ground we may estimate that the effective area with the image might be about 1.9 times that in free space. Mismatches in the system, together with other losses, probably would yield a net value approximately the same as for the free-space condition given above.

For a mismatch at the receiver, producing a standing-wave ratio of Q, the net receiver input power $W_r = 4WQ/(1 + Q)^2$.

41. Double-dipole Antenna. Two dipoles end to end and connected to the two wires of a balanced feeder provide a very useful antenna for simple directive applications. The dipoles may be of arbitrary length, but for fixed frequency working are usually $\lambda/2$, or for greater directivity, $5\lambda/8$. The currents in both dipoles are then cophased, producing maximum radiation normal to the antenna. Its vertical pattern in that plane is the same as a single dipole.

The gain of Fig. 33*a* is 1.8 db over a single $\lambda/2$ dipole, and the gain for Fig. 33*b* is 2.8 db. In this figure only one half of the pattern is shown.

When series-fed at the center and allowing standing waves on the feeder with suitable impedance-matching circuits for coupling to the radio equipment, this antenna will accommodate a wide range of working frequencies while directing the maximum field along the normal to the antenna. Each side is made $5\lambda/8$ at the highest working frequency. As the frequency is lowered, the electrical length of the radiator decreases, the radiation resistance decreases, and the pattern broadens in azimuth. Where a single antenna must be used for several working frequencies, accepting the random variations in the vertical pattern as a compromise, this design is very useful.

42. The Lazy-H Antenna. This system is a combination of two double dipoles (Fig. 34) one above the other, separated $\lambda/2$ and voltage-fed. The horizontal pattern is the same as for a single double dipole, but the vertical directivity is increased by the use of the second set. This, in combination with the effects of height, provides a variety of useful beams which can be focused in the direction and vertical angle required by the propagation conditions. The currents in all four dipoles are cophased by the transposition in the feeder between the upper and the lower radiators. The

[1] See references 40 and 279.

vertical pattern normal to the radiators when over perfectly conducting earth is

$$F(\phi) = \cos\,(H^\circ \sin \phi + 90^\circ) + \cos\,[(H^\circ + 180^\circ) \sin \phi + 90^\circ]$$

where H = electrical height, deg.

Fig. 33. Double dipoles and approximate patterns.

Fig. 34. Lazy-H antenna.

This is the arithmetic sum of the vertical pattern for the upper radiator, added, angle for angle, with the pattern for its image. In performing this addition, the polarity of the field of each lobe must be correctly included.

In free space the gain of this array is 3.95 db over a single dipole. When a plane reflecting screen is added to make it unidirectional, spaced 0.1λ parallel to the antenna, the gain can be increased by another 4 to 6 db depending upon the area of the screen.

The same principle can be used for a vertical-H array for vertically polarized transmissions.

Taking into account the actual ground reflectivity, the vertical pattern is best expressed by

$$F(\phi) = 1 + e^{jS \sin \phi} - Ke^{j(S+2H) \sin \phi} - Ke^{j(2S+2H) \sin \phi}$$

where S = electrical spacing between upper and lower radiators
H = electrical height above ground of lower radiator
K = amplitude of reflection coefficient of ground

Unidirectional transmission is obtained when a close-spaced reflector grid is placed on the blind side of the system.

43. Vertical Antenna. Vertical $\lambda/2$ dipoles are not much used for h-f services. In the past, most uses of vertical dipoles have been where the uniform omnidirectional radiation property was of prevailing importance, but quadrant antennas may be used for this purpose. Vertical polarization for short and medium distances from low altitudes and over sea water is often required, and the vertical $\lambda/2$ dipole can then be employed advantageously. A loaded vertical antenna less than $\lambda/2$ is frequently used as a vehicular antenna[1] (see Sec. 19, whip antennas).

Its vertical field-intensity pattern varies with its height. The pattern function for a $\lambda/2$ vertical dipole over perfectly conducting earth is

$$F(\theta) = \frac{\cos (90^\circ \cos \theta) \cos (H \cos \theta)}{\sin \theta}$$

where θ is the angle with respect to the antenna and H is the height in electrical degrees of the current maximum, above ground. At h.f. the complex reflection coefficient from earth or water causes a mutilation of this ideal pattern. The very low angle intensity is greatly reduced, especially along the ground. This emphasizes another important consideration with all vertical h-f and v-h-f radiators—the nature of the earth out to considerable distances from the radiating system. The performance of identical antennas in different locations may vary greatly owing to earth differences. In the immediate vicinity of the radiator, this can be beneficially controlled by the use of a radial system of elevated, surface, or lightly buried ground wires having a length of the order of $\lambda/2$. The greater the distance between the lower end of the radiator and the earth, the smaller the importance of the wire system in reducing system losses, and the greater should be its length if used.

44. Feeding a Vertical Dipole. The most convenient method of feeding a vertical $\lambda/2$ dipole is at its lower end which is a voltage antinode. A voltage feed method is therefore necessary. If a balanced feeder is used, a line arrangement of transforming from balanced to single-end and to high standing-wave potential is required. One method is to run the open-wire balanced line past the antenna $\lambda/4$ and short-circuit it. The antenna is then connected to one side of this line $\lambda/4$ from the short circuit. This sets up a very strong standing wave in the feeder and equalizes the unbalance introduced by feeding from one wire only. Some distance toward the transmitter the feeder is terminated with a stub line inductance or coupled matching section. In some cases the system can be fed without impedance matching if line losses can be tolerated and the input impedance is of a value that can be accommodated by the transmitter coupling. If the antenna is elevated, the feeder wires rise to the height of the lower end of the antenna and one wire connects to the antenna end and the other remains open-circuit. The unbalance caused in the feeder can be corrected at some convenient point and terminated at another point. Center feed for a vertical dipole is mechanically difficult, but when used, the feeder should be as nearly normal to the antenna as realizable to avoid large unbalance of the feeder. In general, end feed is preferred.

Vertical antennas for direction-finding stations are usually shorter than $\lambda/2$ at the highest frequency, are rigidly held in place near ground, and have extensive radial

[1] See reference 3, p. 282.

ground-wire systems for each antenna, so as to minimize wave tilt in the vertically polarized field component used for direction finding.

45. Folded-dipole Antenna.[1] This radiator is virtually a $\lambda/2$ dipole, a cage, and an impedance transformer built into one unit. H-f types made of wires are exemplified in Fig. 35. As a radiator it functions like an ordinary dipole, giving the same patterns, but with transformed radiation resistance and reactance at the feed point. Greater cross-sectional area due to folding gives it wider band-width characteristics for a given deviation in impedance, and reduces potential gradients and end potentials, for a given input power. Impedance transformations due to mutual impedances between the wires and due to their lengths make it possible to provide exact match with ordinary open-wire balanced feeders. A variety of combinations of wire sizes, spacings, lengths, and foldings are possible for specific applications, and single-end as well as balanced versions can be designed for h-f, v-h-f, and u-h-f uses.[2] For a two-wire folded dipole the factor is 4, and for the three-wire system it is 9 times the radiation resistance for a single-wire dipole. The antenna impedance can be made to match the feeder by choice of spacing and wire size, height, and length. In general, for a folded $\lambda/2$ dipole of N wires of equal diameter, the terminal resistance is approximately $70N^2$ ohms.

FIG. 35. Folded-dipole antennas. Resistances shown are due to height above ground.

The stepup impedance ratio U of a two-conductor folded $\lambda/2$ dipole antenna is given as a function of the conductor radius and spacing by[3]

$$U = (1 + a)^2$$

where

$$a = \frac{\cosh^{-1}\left(\frac{d^2 + \rho_1^2 - \rho_2^2}{2d\rho_1}\right)}{\cosh^{-1}\left(\frac{d^2 - \rho_1^2 + \rho_2^2}{2d\rho_2}\right)}$$

d = center-to-center spacing

ρ_1 and ρ_2 = radii of the two conductors ($\rho_2 \geqq \rho_1$)

[1] See reference 265.
[2] See references 212, 223, and 3.
[3] See reference 226.

The feed point is located at the center of the smaller-diameter conductor.

A three-wire cage 10 in. in diameter has a band width of ±7 per cent from optimum for a maximum of 2:1 standing-wave ratio on the feeder.

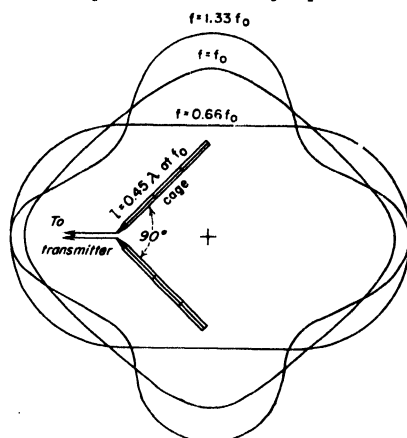
46. Quadrant Antenna (Fig. 36). The quadrant antenna[1] consists of two radiators lying in a horizontal plane and mutually perpendicular from a common apex. Each radiator has a nominal length of 0.45λ for single-frequency use, or for the mid-band frequency when designed to obtain wide-band characteristics of approximately 2:1. Its principal feature as a radiator is its omnidirectional radiation pattern for horizontally polarized fields, as distinguished from a straight antenna. The horizontal field-intensity distribution for the latter varies approximately as the cosine of the angle to the wire, whereas with the quadrant antenna the pattern is virtually square with broadly rounded corners. The wide-band arrangement is often desirable either for eliminating antenna tuning over a 2:1 range or for the transmission of signals requiring great band width. At the extremities of the 2:1 frequency band, the pattern changes from an oval with major axis along the bisector of the quadrant and having an eccentricity of about 1.8 (lowest frequency) through the near-square pattern at mid-frequency to a quadrilobe pattern with a maximum to minimum ratio of about 1.5 at the highest frequency.

Fig. 36. Approximate horizontal field-intensity patterns for wide-band cage quadrant antenna at high, mid-, and low frequencies.

The broad-band characteristics are derived from the use of large-diameter cages having low characteristic impedance. For single-frequency operation, a single-wire design is used. The vertical pattern is determined by the height above ground just as for a straight dipole.

The quadrant antenna is fed at its apex by a balanced transmission line having characteristic impedance between 400 and 600 ohms, depending upon particular designs. Four such antennas, properly proportioned, can be used to cover the h-f band in 2:1 steps, and these can all be assembled on four supports. Identical arrangements may be used for transmitting and receiving.

47. Extensive Dipole Arrays. The H antenna is the basic form for flat arrays of dipoles which can be extended as far as desired in height and width to achieve a high degree of radiation control in vertical and horizontal planes. In the same manner, arrays can be made with dipoles in depth in any degree, in what are called *cubical arrays*. In practice, cubical arrays have found almost no use because of the complications of construction. Unidirectional patterns have been mainly obtained by placing a duplicate sheet of dipoles behind the main sheet and exciting the reflector parasitically or directly.[2] The adjustments of the reflector elements differ somewhat for the two conditions. Another, simpler method is to use a close-spaced neutral reflector sheet of area somewhat larger than the radiator sheet made up of wires spaced 0.05 to 0.1λ parallel to the radiating dipoles. The closeness of spacing of the reflector wires primarily determines the magnitude of the backward radiation, and the forward radiation is but slightly increased for wire spacings smaller than 0.2λ.

Through the years a great number of distinctive beam-antenna designs have appeared and disappeared. It would be possible to describe numerous ingenious forms developed in the great era of h-f growth from 1923 to 1932. In the interim,

[1] See reference 237.
[2] See references 188, 197, 198, 207, 218, and 232.

economics and the virtual abandonment of vertical polarization for long-distance h-f communications took their toll of all but a very few basic forms which are considered here in some detail because of their importance, technically and economically. It is interesting to note that evolution has shown the absurdity of paying a high price for an extra decibel of gain when the medium of propagation varies many hundreds to one. This variation in the medium has also obscured the very poor performance of many antenna designs. Multipath transmission places a limit on the speed of keying and on speech mutilation on very long circuits. Greater control of the vertical patterns for transmitting and receiving antennas can effect improvements by angular discrimination between various orders of hops with various delays. Owing to multipath delays, the working speed decreases as circuit lengths are increased. On very

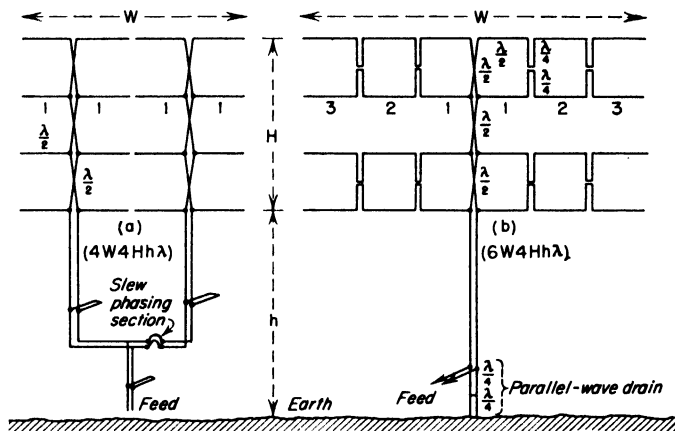



FIG. 37. Arrays of horizontal half-wave dipoles and typical feeder connections.

long circuits, a fundamental limitation exists where the two or three lowest orders of hops come within angles which overlap those due to variation in layer height, and against which angular discrimination is impossible. In such cases the only solution seems to be to use intermediate automatic relays.

The flat array of horizontal dipoles (the *pine-tree* or *Koomans array*) is a basic type which provides independent control of vertical and horizontal patterns.[1] The V and rhombic types have these two properties interrelated.

4W4H and 6W4H arrays are shown schematically in Fig. 37. Dipoles are each $\lambda/2$ and spaced $\lambda/2$. Diagram *a* shows two feedlines, the array being two sets of four cophased double dipoles. This feeding method gives equal currents in all dipoles at one level, and there is a slight reduction in currents vertically due to attenuation along the feeders. At *b* is a common-feeder system with three dipoles in series, cophased by the $\lambda/4$ phasing sections between colinear dipoles. Cophasing between levels is by transposition of the feeder. Identical arrays could be used as reflectors, or a neutral sheet reflector employed.

In a multifeeder system as in *a*, the two sections of the array can be mutually dephased by a small amount to slew the beam a few degrees for different circuits with small angular path difference.

The horizontal pattern function for an array of *n* colinear cophased dipoles arranged in *m* symmetrical dipole pairs from the feeder is given approximately by

$$F(\theta) = \frac{\cos(90^\circ \cos\theta)}{\sin\theta} \sum_{m=1}^{n/2} I_m\{\cos[(2m-1)90^\circ \cos\theta]\}$$

[1] This is only approximately true because of the mutual coupling between elements. However, in most practical arrays it is a good approximation. See reference 51.

where the coefficient is the pattern for a single dipole, and each term in the series is the pattern for an isotropic source located at the center of each dipole pair symmetrically from center (feeder) and having an intensity proportional to its loop current. Azimuth angle θ is measured from the line of the radiators. This pattern can be expressed in many other forms, but this one permits the use of current-grading coefficients as well as the usual equality of currents.

The vertical pattern for an array of k vertical elements can similarly be obtained from the following equation, where each level of radiators forms a pair with its image but with image current reversed.

$$F(\phi) = I_1 \cos (H \sin \phi + 90°) + I_2 \cos [(H + 180°) \sin \phi + 90°] + \cdots + I_k \cos \{[H + (k - 1)180°] \sin \phi + 90°\}$$

At an elevation angle ϕ, with a height H deg between ground and the lowest radiator with current amplitude I_1, the first term will be the pattern for this first level. Other levels have patterns according to the successive terms at successive heights differing by $\lambda/2$. If other than $\lambda/2$ vertical spacing is used, the actual spacing is substituted for 180 deg in this equation. This equation is adapted to the use of vertical current grading by assigning proper values to I_k. Actually to grade the currents, the feeder line must act as an impedance transformer while retaining its proper phase relations at the various radiator levels.

For identical currents in all dipole elements, the maximum field-intensity gain due to directivity is proportional ideally to the square root of the number of dipoles used, in both radiator and reflector curtains. With neutral reflectors, the radiator images are counted. A number of factors cause an appreciable impairment of this ideal gain, among them ground losses, insulation losses, imperfect phase differences, reflections from discontinuities or irregularities, conductor losses, losses in associated rigging and supports, etc. A well-constructed system may be 2 to 6 db below expectation depending upon size, frequency, and location.

In constructing such a system, all conductors are cut from 5 to 10 per cent shorter than theoretical length based on a velocity of propagation equal to that in free space, to allow for effects of corner junctions, insulator capacitance, and end effect.

48. Tuning Dipole Arrays. The tuning of an array using a tuned reflecting sheet is very difficult because of mutual impedance relationships, which cause very strong interactions between radiators and reflectors. When the reflector is excited parasitically, it must be quite close to the radiator sheet to have best effectiveness. The complex nature of mutual impedances causes the spacing and adjustments to be different for maximum forward radiation as compared with minimum backward radiation. To reduce this tuning difficulty, some systems have used a radiator-reflector spacing of $5\lambda/4$ with each sheet fed directly but with a quadrature phase relationship. The objection to the latter is that additional supports are required for the reflector, whereas the close spacing permits the use of the same supports with crossarms or the use of spreaders. The simplest system is to use a neutral reflecting screen. Tuning is then confined to the one active sheet of radiators, and the spacing is diminished. If a reversible antenna is required, another sheet of radiators can be placed on the opposite side of the reflecting screen, all using one set of supports. If both radiating systems are tuned to the same frequency, a detuning stub is switched into the idle system.

49. Dipole Array Feeders. Feeders are always kept normal to radiators to maintain correct balance, and impedance matching of the feeder is done only at one point outside the antenna system, provided all internal conditions are correct. In multiple-feeder systems, each main feeder can be terminated before being joined to the main power trunk, which in turn is matched near this junction. Care must be observed to ensure that identical conductor lengths are included in both sides of the feeder when making bends, through switches, etc. If there is radiation coupling between radiators and feeder which sets up a parallel wave to ground, the nodal points in the feeder standing-wave pattern will be displaced mutually on the two sides of the feeder. To eliminate the parallel wave, one end of a $\lambda/2$ stub can be connected across the feeder

at some point and to a good ground at the other end, with a short circuit across the stub at $\lambda/4$ from the feeder junction. This stub is conveniently inserted vertically between the lower radiators and ground, as shown in Fig. 37*b*. The main feeder can then be terminated somewhere near the end with another stub or coupled section. On a very long feeder, the cumulative effect of line irregularities caused by insulators and bends may require another matching stub nearer the transmitter.

Dipole arrays of this type, made of single wires, will transmit band widths of the order of ± 2 per cent of optimum. To broaden this band slightly, all conductors in

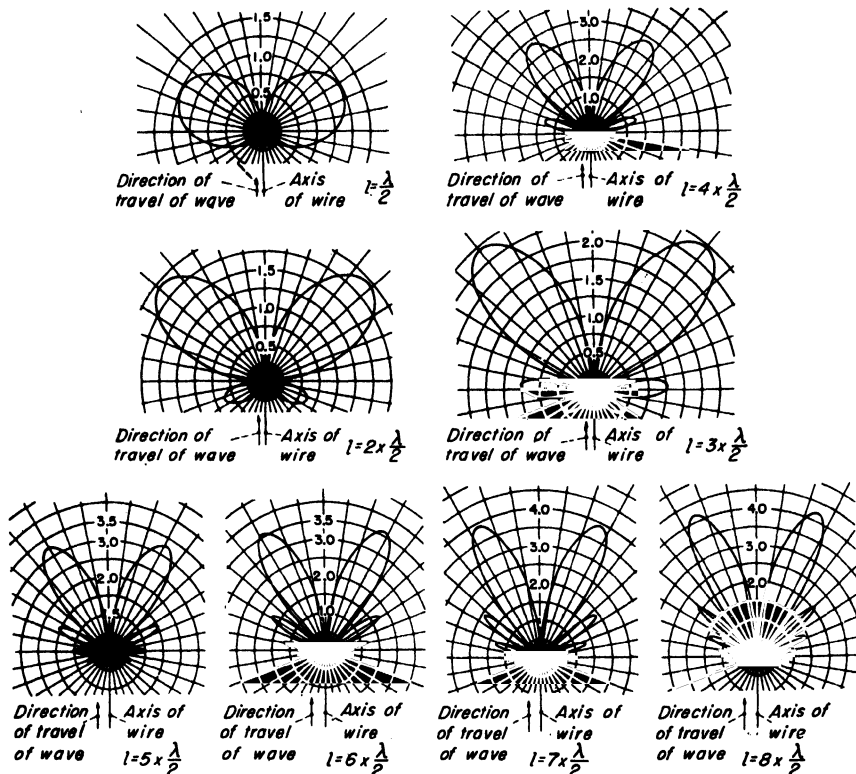



FIG. 38. Field strengths due to straight wire carrying a traveling wave.

radiators and feeders must have low characteristic impedances, most conveniently accomplished by using double-wire hairpin radiators and four-wire feeders.

An array 2W4H (eight dipoles), with lowest radiators 1λ above ground, made of 0.100-in.-diam wire, and with 8-in. porcelain insulators with $3\frac{1}{2}$-in.-diam end fittings (insulator capacitance 1 $\mu\mu$f), has a wire length at resonance of about 0.45λ per dipole at 18 Mc. With 100-kw carrier input, the end potential on each dipole is 7,500 volts rms. A 3,000-ft 580-ohm two-wire balanced feeder used with this antenna had a transmission efficiency of 67 per cent.

Guyed supports for h-f antennas should be removed as far as possible from the radiator fields. It must be remembered that the radiation pattern customarily referred to is that at great distance, and that very near an array there are likely to be strong fields in any direction from the radiators. The use of breakup insulators in guys is determined by their natural period. Sometimes an insulated guy will resonate with the effect of insulator capacitance between sections when an uninsulated guy would be neutral. Few or no breakup insulators are preferred when guys can be

removed from strong fields and be of nonresonant lengths. The current flowing in any conductor is the induced potential over the impedance. In a strong field, the impedance of a guy wire or support must somehow be made as high as possible to suppress parasitic currents.

50. Traveling Waves on Long-wire Antennas. A long-wire antenna carrying a traveling wave without attenuation has the normalized radiation-pattern function

$$F(\theta) = \sin\theta \frac{\sin\left[\frac{\pi l}{\lambda_0}(c/v - \cos\theta)\right]}{\frac{\pi l}{\lambda_0}(c/v - \cos\theta)}$$

where θ = polar angle measured from the direction of the wire
c = free-space velocity
v = velocity of propagation on the wire
l/λ_0 = antenna length in free-space wavelengths

For most thin linear antennas $c/v = 1.0$; however, for other geometries c/v can be made greater than or less than unity. The first term in the above equation is the pattern function of an element of the wire, and the second term is the array factor for a continuous distribution in the form of $(\sin x)/x$.

Figure 38 shows a series of polar diagrams which are cross sections of the solids of revolution surrounding wires of different lengths. The intrinsic characteristic of diminishing amplitudes for successive lobes, following the main one closest to the direction of the wire, is a valuable property in directive applications. The four sides of a rhombic antenna are each radiators of this type, with the geometry arranged to add the main lobes and further suppress the minor lobes by wave interference. The fact of attenuation present in practice has but slight influence on the major lobe, and its principal effect is to fill in the nulls between lobes. A rigorous treatment of this type with attenuation and effects of ground has been published by Cafferata.[1]

51. Standing Waves on Long-wire Antennas. In the same manner, a long wire in free space m half waves long with the standing-wave distribution of current shown in Fig. 39 produces field distributions following these equations.[2]

For m odd,

$$F(\theta) = \frac{\cos\left(\frac{m\pi}{2}\cos\theta\right)}{\sin\theta}$$

Nulls occur at

$$\cos\theta = \frac{1}{m}, \frac{3}{m}, \ldots, \frac{m}{m}$$

Maximums occur at

$$\frac{m\pi}{2}\tan\theta\sin\theta = \cot\left(\frac{m\pi}{2}\cos\theta\right)$$

For m even,

$$F(\theta) = \frac{\sin\left(\frac{m\pi}{2}\cos\theta\right)}{\sin\theta}$$

Nulls occur at

$$\cos\theta = 0, \frac{2}{m}, \frac{4}{m}, \ldots, \frac{m}{m}$$

Maximums occur at

$$\frac{m\pi}{2}\tan\theta\sin\theta = \tan\left(\frac{m\pi}{2}\cos\theta\right)$$

These also neglect attenuation. The same general phenomenon as noted in the previous paragraph is present, except for the symmetry of the pattern. Figure 39 exhibits

[1] See reference 195.
[2] See references 16 and 197.

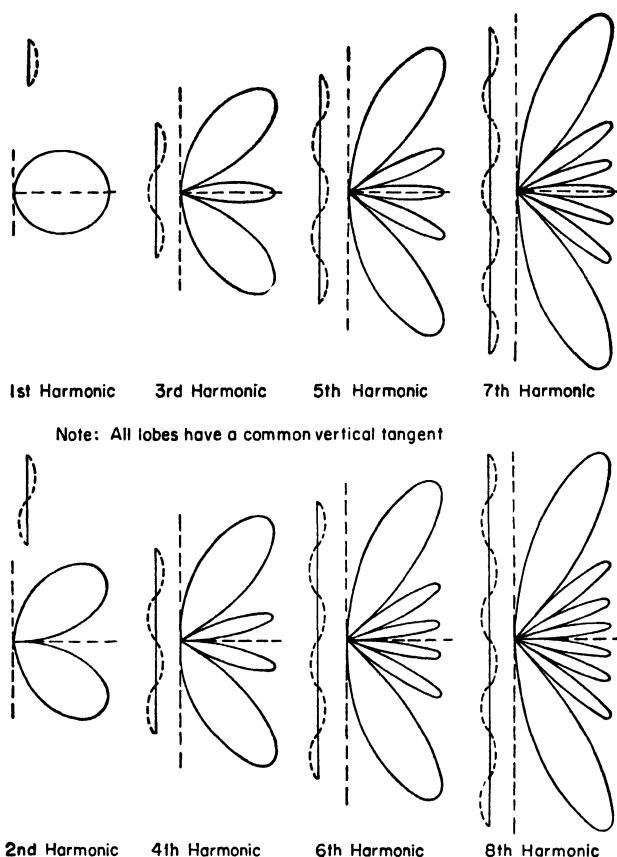



FIG. 39. Polar diagrams of relative field-strength distribution for straight-wire antennas in free space with standing waves of current as shown, corresponding to the operation at various harmonics of the fundamental frequency of the antenna.

a series of patterns for medium lengths, and Fig. 40 compares directly a $7\frac{3}{4}\lambda$ and an 8λ pattern. This current distribution can be obtained only if the antenna is either "end-fed" or energized at a current maximum (shown in Fig. 39) by a balanced two-wire transmission line.

The free-space radiation-pattern function of a center-fed linear antenna of length l with a sinusoidal current distribution is

$$F(\theta) = \frac{\cos\left(\dfrac{\pi l}{\lambda_0}\cos\theta\right) - \cos\dfrac{\pi l}{\lambda_0}}{\sin\theta}$$

where θ is measured from the axis of the antenna.

The current distribution in this type of antenna is essentially sinusoidal if the antenna is thin (diameter less than $\lambda/100$).[1]

52. Horizontal Rhombic Antennas.[2] This basic type of antenna has won a permanent place in the h-f field. It is a traveling-wave system suitable for transmission and

[1] See references 9 and 11.
[2] References 192, 193, 194, 195, 197, 201, 214, and 222.

reception. It has singular merit for its broad frequency response. It is relatively inexpensive to construct and maintain and requires no tuning. Its main disadvantage is that it requires a large plot of land. A minor disadvantage is that its horizontal and vertical patterns are dependent one on the other, which sometimes imposes limitations on its effectiveness. The sharper the horizontal beam, the lower the angle of the vertical beam, and it is impossible to obtain high-angle radiation except at very low gain and very broad horizontal pattern. Another disadvantage is the need to dissipate a considerable proportion of the total input power uselessly in the terminating resistance to obtain aperiodicity and unidirectionality. Except for the very high cost of r-f energy per kilowatthour, this is not generally regarded as a serious deficiency in view of its other merits and in relation to the fact that the propagation medium varies normally so much in proportion as to obscure this loss.

Envelope parallel to wire

Radiation pattern $7\frac{3}{4}\lambda$ wire (sinusoidal current distribution)

Radiation pattern 8λ wire (sinusoidal current distribution)

FIG. 40. Comparison of radiation patterns for standing-wave current distribution of $7\frac{3}{4}\lambda$ and 8λ wires.

In transmitting, the potentials in the system are lower than for resonant systems. For the usual three-wire antenna with characteristic impedance of approximately 600 ohms, the maximum potential (at the input end) is approximately 7,750 volts rms for 100 kw carrier input. This is a desirable characteristic for high-altitude applications.

Rhombic horizontal antenna patterns are complicated by the skew geometry of the array and by the attenuated traveling waves. The main beam, usually the center of design interest, can be handled with adequate exactness by neglecting the attenuations, the complex dielectric properties of the ground, and the vertically polarized components of radiation. With these simplifications, the following equation[1] supplies the three-dimensional pattern:

$$F_{\theta,\Delta} = \left[\frac{\cos(\phi - \theta)}{1 - \cos\Delta\sin(\phi - \theta)} + \frac{\cos(\phi + \theta)}{1 - \cos\Delta\sin(\phi + \theta)}\right] \times \sin\left\{\frac{\pi l}{\lambda}[1 - \cos\Delta\sin(\phi + \theta)]\right\}\sin\left\{\frac{\pi l}{\lambda}[1 - \cos\Delta\sin(\phi - \theta)]\right\} \times \sin\left\{\frac{2\pi H}{\lambda}\sin\Delta\right\}$$

[1] See reference 17.

where ϕ = one-half side angle of rhombus
θ = azimuth angle with respect to major axis of rhombus
Δ = elevation angle with respect to horizon
l = length of one side of rhombus
H = height of wires above ground

The axial vertical pattern is obtained by placing $\theta = 0$. The horizontal plan pattern through the maximum in the axial vertical pattern is found after substitution of Δ max in this equation and solving for variable θ.

The main beam is horizontally polarized and is directed along the major axis of the array away from the feeder end. Long arrays have a multitude of smaller beams in all directions in vertical and horizontal angles. Owing to the attenuation of the traveling waves, there are no nulls in the patterns for the individual sides of the rhombus except in the axis of the wire, which is blanked by the pattern for the other three sides. There is no vertical or horizontal angle at which there is zero radiation, even though relatively very small in magnitude.

The antenna may be regarded as a transmission line which diverges from the feeder until it reaches an electrically large separation, then reconverges to the forward end virtually to a point. As the wires diverge, energy escapes from the system because there is no canceling effect from the opposite sides such as occurs in a close-spaced balanced transmission line. This divergence, however, is accompanied by a change in characteristic impedance unless means are taken to equalize it in relation to itself and to ground. An irregularity is introduced at the side corner so that the degree of suppression of standing waves on the terminal side is usually different from that on the feeder side. When its characteristic impedance is equalized by a spreading of three wires on each side, a more uniform match occurs throughout the system. In the case of receiving, this uniformity contributes to suppression of noise pickup from the rear. It is reported from comparative observations that a multiwire rhombic antenna is appreciably less susceptible to precipitation static interference than a single-wire system.

The three fundamental variables that control the radiation pattern are the electrical length of one side, the electrical height above ground, and the side angle at the bend. In all references, one-half of this side angle ϕ is employed as a parameter. We retain the established nomenclature for rhombic antennas herewith and assign ϕ to the half-side angle, θ to the azimuth angle with respect to the forward axis of the array, and Δ to the elevation angle. By various combinations of these factors, the main beam is directed in elevation and the beam width controlled in azimuth. At the same time the minor lobes are suppressed or enlarged in various directions so that the gain of the main beam can be slightly optimized by fine adjustment of all three parameters. Data based on an optimized design are presented in Fig. 41.

53. Rhombic Usage. The frequency range over which a rhombus can be used depends upon more than the mere fact of maintaining an impedance match which is satisfactory. Frequencies must be changed to accommodate changes in the propagation medium, and the function of a proper antenna application is to direct the radiant energy where it will be most effective. The desired radiation characteristics for various h-f circuits for various distances, times of day, seasons, and years must be determined for individual applications. Whether or not the use of a single rhombus fits the conditions encountered for the operating frequencies to be used must be investigated in relation to propagation requirements. Multipath transmission due to multiple hops and from varying heights of the different ionosphere layers imposes one of the most serious limitations on transmission speed and fidelity, and the vertical directivity of the transmitting and receiving antennas has much to do with multipath discrimination. Careful engineering of a propagation circuit may require different antennas for different frequencies regardless of the ability of a single antenna to accommodate a wide range of frequencies from an impedance-match standpoint.

In ordinary applications it has been customary to employ one rhombus for all frequencies in one direction, letting the pattern change as it will and accepting random effects in propagation. Where cost is more important than performance, this is the

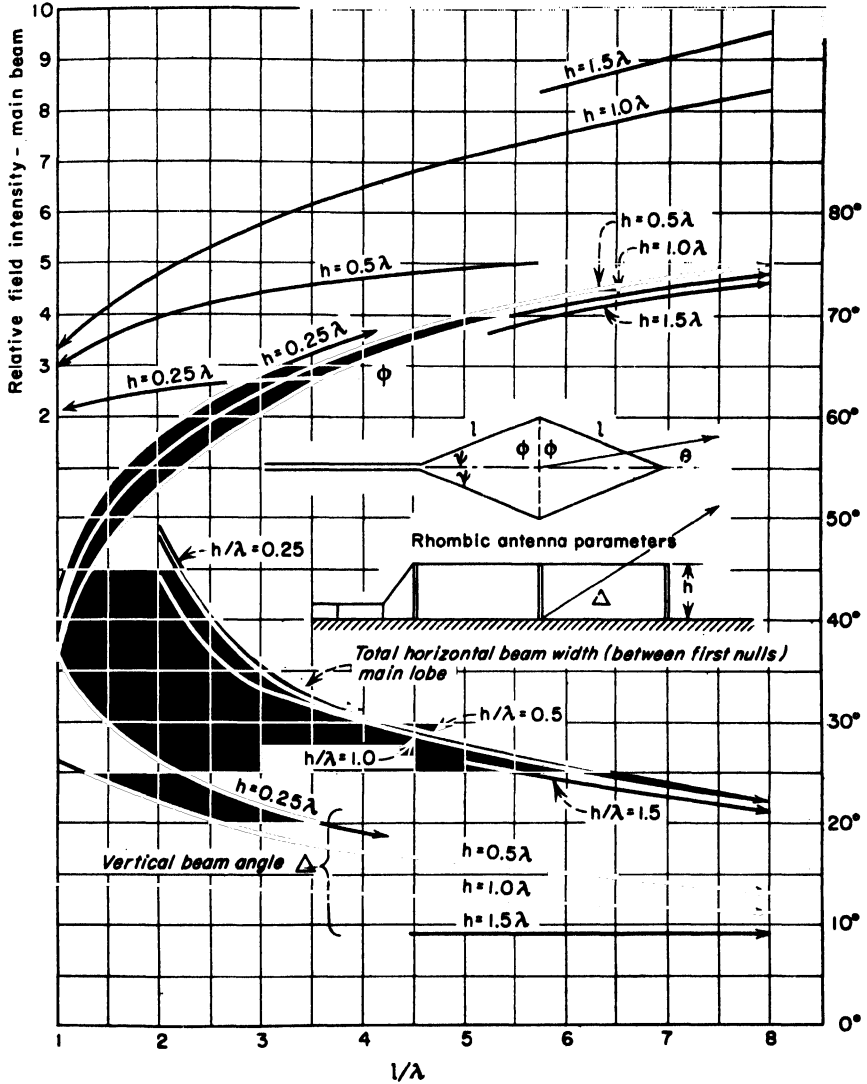



FIG. 41. Rhombic-antenna design data.

best that can be done. In cases where optimum use is made of the propagation medium and where performance is the primary objective, it is often best to use a different antenna specially designed for each frequency for each point-to-point circuit. When this is done, the transmitting antenna need not be aperiodic, and the energy normally thrown away in a dissipative termination can be utilized by conducting it back to the input end via transmission lines which have the correct characteristic impedance, and phased and transformed to reenter the system in parallel with that from the transmitter. This reentrant feed method then improves the efficiency of the system for the operating frequency for which it is designed.

54. Receiving Rhombic Antennas. For receiving, either for single- or multifrequency operation, the antenna is resistance-terminated with a value as close as possible to its average characteristic impedance. Heavy-duty noninductive resistors are used, since small ones are frequently destroyed by induction from lighting. A static drain circuit to ground from the neutral of the termination is desirable to prevent the accumulation of static charges during precipitation. For transmission, where the dissipation is appreciable, high-loss transmission lines are used for terminating the antenna. Lines made of iron or stainless steel are commonly used and are assembled on low supports underneath the antenna, back and forth, to the required length. The

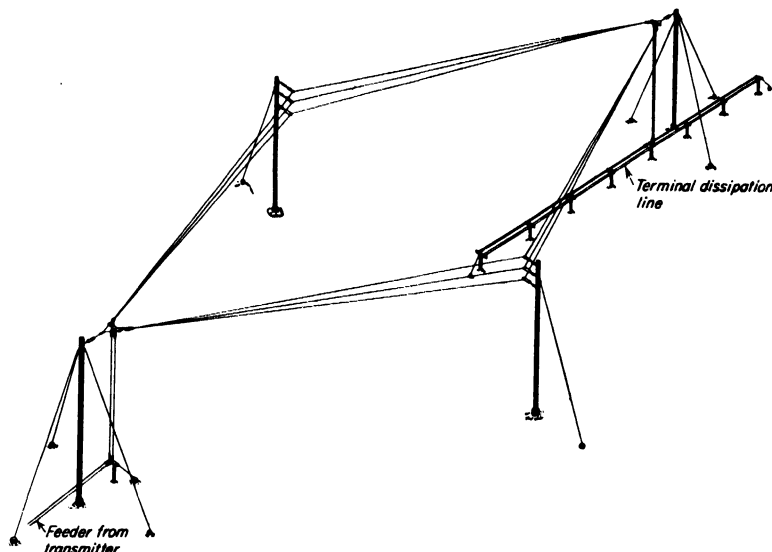



FIG. 42. Horizontal rhombic antenna.

length should be that which will provide at least 20 db attenuation at the lowest working frequency. Such a length is sufficient so that the reaction on the antenna is negligible whether open- or short-circuited. The latter is preferred because it can be directly grounded, thus serving as a static drain circuit also. The characteristic impedance is computed from formulas used for balanced two-wire lines.

The attenuation per unit length is given by the formula

$$\alpha = \frac{54.8 \times 10^{-6} (\sqrt{\rho\mu f}/d)}{Z_0} \quad \text{db/m}$$

where ρ = resistivity, microhm-cm
μ = magnetic permeability
f = frequency, Mc
d = wire diameter, cm
Z_0 = characteristic impedance, ohms

For the line to appear as a virtually pure resistance (phase angle of impedance less than 7 deg) the $\omega L/R$ per unit length at the lowest frequency must be less than 0.1. The inductance per loop meter is

$$L = \frac{Z_0}{3} \times 10^{-8} \quad \text{henrys/m}$$

A typical layout of a three-wire rhombic antenna with dissipation and feeder lines is shown in Fig. 42.

A rhombic antenna is a balanced system, and care must be taken to preserve good balance to ground. If a number of systems are located in the same plot, sufficient spacing must be allowed to maintain low mutual impedances between the nearest sides of adjacent antennas. When two rhombics are used in broadside as a single system, one common support can be used without balance problems, when fed symmetrically. When used with concentric transmission lines, care must also be taken to make a correct balanced-to-single-end transformation.

Fig. 43. Double-fishbone antenna.

55. Fishbone Receiving Antennas.[1] The fishbone antenna (Fig. 43) is an evolution of the wave antenna, adapted to the reception of horizontally polarized waves in the h-f band. Three different designs are used to cover the range 3 to 20 Mc, the differences being in the lengths of the dipoles. Fishbones are used singly, or with two in broadside. A typical system is 312 ft long. A single-bay system requires 10 poles and a two-bay system 16. The essential information on this type of antenna for construction purposes is tabulated below:

Length, doublets, ft	Optimum freq., Mc	Useful range, Mc	Width, 2 bays, ft	Length, total, ft	Best pole height, ft	Useful azimuth angle, deg
34	18	13-22	120	312	60	10
48	14	10-19	148	312	90	10
66	9	3-13	200	312	120	14

The schematic arrangement of the fishbone antenna is shown in Fig. 43. A central transmission line runs horizontally in the direction of the incoming signals. At equal intervals along this line, horizontal doublets are attached which are short enough to be nonresonant within the frequency range of the system, and near enough together to provide a uniform loading of the center line. Each pair of doublets loads the line with parallel capacitance and resistance. To keep the propagation velocity of the line above 90 per cent of free space, each doublet has a series capacitance near the point of attachment. The characteristic impedance of the main feeder is approximately 400 ohms. Two bays in parallel give an impedance of approximately 200 ohms, which can be matched directly into a balanced four-wire line to the receiving station. The end of the feeder toward the transmitting station is terminated in a noninductive resistance equal to the characteristic impedance of the feeder. This makes the system unidirectionally responsive by absorbing energy received from the rear. The fishbone antenna is a high-gain, broad-band, moderate-cost antenna with minimum land requirements. Even though the design is several years old, it remains one of the best receiving antennas for h-f point-to-point communication services.

56. Harmonic Wire Antennas of V Form.[2] Radiating systems using long wires having essentially sine current distributions have been called *harmonic wire antennas*. Several forms have been evolved, of which the two shown are typical (see Figs. 44 and 45). They have moderately high gain and relatively low cost. The main beam,

[1] See reference 189.
[2] See reference 197.

along the axis of the V, is horizontally polarized. The vertical pattern can be controlled to a certain extent by the height of the system above ground, but this is also related to the horizontal pattern in that the sharper the horizontal pattern, the lower the angle of the vertical pattern.

Figure 40 shows the comparison of the field-intensity patterns surrounding harmonic wires having lengths of $7\frac{3}{4}\lambda$ and 8λ. The main lobe nearest the wire axis contains most of the total radiated power. V antennas are arranged to utilize this main lobe from all wires of the system so that they add along the array axis and cancel as much as possible in all other directions. In Fig. 44 the two wires forming each side of each V are cophased, and the two sides of each V are antiphased. The rear V is fed 90 deg in advance of the forward V with equal currents to provide best reflecting action.

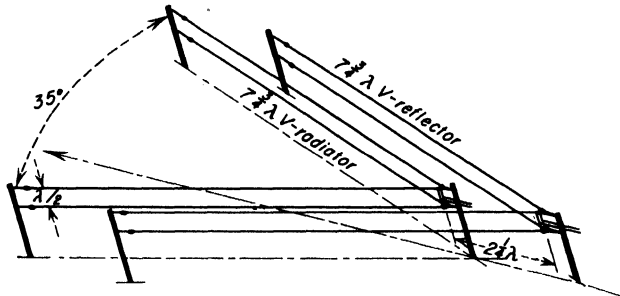



FIG. 44. V antenna with reflector.

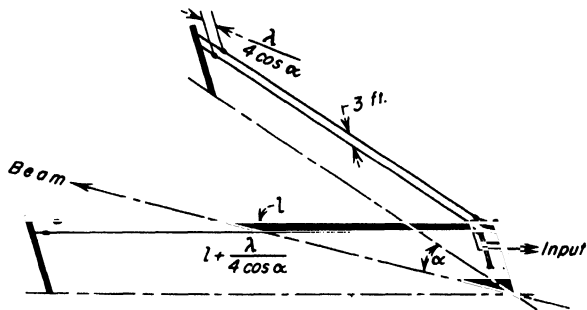



FIG. 45. Unidirectional V antenna.

Figure 45 shows the simplest type of V antenna having unidirectional pattern. The lower wire in each side is longer than the upper by an amount equal to $\lambda/4$ in its projection upon the axis of the V and is excited parasitically from the upper wires, which are energized from the main feeder. Excited in this manner, the two wires on one side are approximately the equivalent of a single wire with a traveling-wave current distribution, producing a unidirectional pattern.

57. Antenna Systems for Ionospheric Forward-scatter Propagation. "Beyond-the-horizon" point-to-point communication is possible at frequencies above the maximum usable frequency (MUF) using a propagation phenomenon called ionospheric forward-scatter propagation. An ionospheric forward-scatter communication system consists of a high-power transmitter supplying a highly directive transmitting antenna oriented to direct its radiated energy toward the ionosphere at a low elevation angle in the direction of the great-circle path to the receiver location, and a highly directive receiving antenna oriented in the direction of the transmitter to receive the energy scattered from the ionosphere.

Scattering is caused by a turbulent region in the lower part of the ionosphere 40 to 200 miles above the earth with the scattering source (the volume illuminated by the transmitting antenna) generally located midway in the path. Since the source is in a turbulent state, the scattering is generally diffuse in nature, complicated by meteoric ionization and other phenomena. The scattering mechanism has a profound influence on the characteristics of the antenna system.

Only a few considerations pertinent to antenna design can be listed here. For details the reader is referred to the literature on the subject.[1]

The most useful range of path lengths for this type of propagation is from about 1,000 to 2,000 km.[2] Received signal intensity decreases rapidly with decreasing distance below about 1,000 km. The frequency range of greatest usefulness is from 25 to 60 Mc. Frequencies below about 25 Mc are subject to interference due to sporadic ionospheric propagation. In general, the lower frequencies in this range have been found to be best.

As stated above, high transmitter powers must be used because of the scattering mechanism. For simple communication systems, 5 kw of transmitter power is usually adequate at frequencies near 35 Mc using rhombic antennas. If the receiver used has a sufficiently low noise figure (an NF of 2 is adequate), external noise will be the limiting factor in this frequency band.

Multipath propagation from strong off-path reflections due to meteoric ionization, when it occurs, is a factor which reduces the propagation-path band width. The antenna side lobes should be sufficiently low to provide adequate suppression of off-path signals. Antenna beam widths should be narrow in azimuth for the same reason; however, there are other beam-width considerations stated later. Propagation-path band width will be limited by either antenna beam width or the natural beam width associated with the scattering process. Band widths of 50 kc can usually be obtained.

The following is a table of desirable antenna characteristics.

Desirable Antenna Characteristics*

Characteristic	Value
Horizontal beam width	8 deg or less
Vertical beam width including ground effects	6 deg or less
Maximum side-lobe level	40 db below main beam
Radiation efficiency	90% or greater
Horizontal orientation of beam	Normally on great-circle bearing between transmitter and receiver
Vertical orientation of beam	85 km above mid-point of great-circle path
Provision for reducing directivity or for varying direction of beam	For use during periods when scattering is not homogeneous and during periods when scattering meteoric components will provide higher signal-to-noise ratio
Band width over which characteristics are to be maintained	200 kc

* See reference 186.

The use of spaced-antenna receiving diversity is desirable. The diversity antennas should be placed in a line normal to the path with not less than 10 wavelengths spacing.

In general, horizontal polarization is preferred because the scattering losses are somewhat less than for vertical polarization. In addition, the reflection characteristics of the ground are more desirable, and, too, it helps to discriminate against man-made noise.

The *realized gain* of a directive antenna (defined as the ratio of signal intensity observed when the antenna is used to that observed when a reference or comparison antenna is used during the same period of time) is a function of the directivity of the antenna used at the opposite terminal and is influenced by diurnal variations in the propagating mechanism. Realized gains both greater than and less than *plain-wave gains* have been experienced. Refraction in the troposphere also affects realized gains, especially at low angles of radiation.

[1] October, 1955, *Proc. IRE* is devoted to ionospheric and tropospheric forward-scatter propagation. See reference 241.

[2] See reference 186.

In siting the antenna for low-angle radiation, the region in which ground reflections occur, which may be quite long for very low angles, should be flat to within a quarter of the antenna height if the 6-db increased gain is to be realized due to ground reflections.

To date, long rhombics 25 to 50 wavelengths on a side, arrays of smaller rhombics, large corner-reflector antennas, broadside arrays of corner reflectors, broadside arrays of dipoles, and arrays of Yagi antennas have been used in this service.

V-H-F ANTENNAS (30 TO 300 MC) AND U-H-F ANTENNAS (300 TO 3,000 MC)

Developments in communications, f-m and television broadcasting, radar, navigational aids, and many other applications have given birth to an endless variety of antennas during recent years. This treatise cannot possibly include reference to all of them though certain basic types must be included. From these, other types can be understood. The mechanical design of antennas in these bands departs radically from those for lower frequencies because supported wires usually are replaced by rigid members, many insulators are eliminated, metallic line sections are used as insulators and in other ways. Some antennas must respond uniformly to very large band widths; some must have very high directivity; others must have switchable patterns, rotating or waving patterns, and other special properties. Since such antennas are often located in high places subject to severe weather, designs must be extremely robust. Most forms are prefabricated before erection. Electrically some are very complicated, with ingenious forms of radiators and feeding systems. Some use sheet distributions of currents instead of linear distributions, magnetic instead of electric dipoles, and many use flat, angular, or parabolic reflecting screens. In the higher-frequency portion of the u-h-f band, wave-guide feeders are frequently used instead of ordinary transmission lines. All the antenna technology found in all the lower frequencies is employed in these bands together with many other principles peculiar to these bands.

58. Electric Dipoles. A description follows of certain basic radiator types used alone or in antenna arrays (see Fig. 46).

Straight Electric Dipole (Fig. 46*A*). A straight rod or tube of length approximately $\lambda/2$, mounted with a metallic rod attached at the mid-point (zero potential) and normal to it. To increase its band width, its diameter is made relatively great, or it may be ellipsoidal instead of cylindrical in extreme cases.[1] It can be shunt-fed with a balanced transmission line.

Figure 46*B* is a center-fed dipole supported by a $\lambda/4$ transformer. One side of this support acts as a continuation of the concentric feeder and, where the inner conductor emerges at the center of the radiator, it is connected to the other side. This arrangement uses the $\lambda/4$ supporting insulator also as a single-end-to-push-pull transformer to provide the required balanced feed potential at the center of the radiator.

Figure 46*C* is a form where both sides of the $\lambda/4$ support insulator act also as concentric feeders. In this case, charges flow out from the inside of the outer conductors of the feeders and propagate along the outsides of the transformer and radiator where they set up standing waves in the usual way. The inner conductors for the two feeders are connected together.

Figure 46*D* shows a method of dipole construction by which the impedance at the center point of feed can be made resistive in any ordinary amount required for correct balanced feeder matching.[2] The $\lambda/4$ supporting transformer can be shunt-fed from a balanced line, as at points *B*, or the radiator can be excited directly when the feeder is attached at points *A*. The charges that flow on the outside of the tubular conductors set up the usual standing wave of current. The length is made somewhat less than $\lambda/4$ each way from center so that, at the feed point, the impedance appears as a resistance in series with a capacitive reactance. Currents propagated inside the tubes, by virtue of the concentric conductor in the inside acting as a transmission line

[1] See reference 294.
[2] See reference 259.

(and also as a mechanical support for the two halves of the radiator), are reflected from short-circuiting disks, which make the inner lines appear as inductive reactances. When these are adjusted so that this reactance forms a parallel-resonant combination with the external impedance of the radiator itself, the feed impedance is made resistive. By proportioning the external length and the internal short-circuit position, the feed-point resistance can be adjusted over a wide range to a desired value.

Figure 46*E* uses somewhat the same principle as the previous form, except that the inductive tuning to correct the power factor of the radiator portion is obtained by placing a short circuit across the $\lambda/4$ support transformer at the correct position. It can be fed with a balanced feeder at points *A* or *B*.

These five forms are excellent examples of multiple utilization of all parts of the system for electrical and mechanical purposes. Those using the $\lambda/4$ transformers for supports are ordinarily attached to reflecting sheets or to some structural members for an array of such radiators. They lend themselves to virtually all-metal welded construction with a minimum of insulation.

Figure 46*F* is a form frequently used to obtain a vertical radiator which terminates a concentric feeder.[1] The outer conductor is, in principle, turned back $\lambda/4$ while the center conductor continues outward for $\lambda/4$. This gives a vertical dipole in which the image charges from the inside of the sheath balance those flowing upward in the projection of the inner conductor. The skirt forming the lower half of the dipole is insulated from the sheath of the feeder and suppresses sheath currents on the feeder below it. In practice, however, there is considerable coupling to the feeder from the skirt, and it should be taken into account in the design. Since the feed-point impedance of such a radiator is very nearly that common to many types of coaxial line, a satisfactory impedance match and single-end-to-balanced transformation are achieved simultaneously.

Figure 46*G* is a radiator less than $\lambda/4$ long that projects from a short coaxial stub section which is short-circuited near its lower end.[2] The antenna presents an impedance of resistance and capacitive reactance which is tuned to a resistive condition at the feed point at the upper end of this stub, acting as a tuning inductance. Two or four horizontal $\lambda/4$ rods project outward from the lip of the concentric stub, which is attached to a mast. The impedance at the lip is adjusted by radiator length and stub length to be correct to match a concentric feeder attached to the mast. The radial rods carry the image charges for the radiator portion and by virtue of their $\lambda/4$ length bring a potential node at the transmission-line sheath. This suppresses currents on the sheath below. In this way a perfect impedance match can be made, and the assembly is compact and rugged. A folded-monopole antenna can be fed in a similar manner.

The forms in Fig. 46*H* and *J* do the same thing in slightly different ways. In the former, the suppressor section, in the form of a cylindrical skirt connected to the outer sheath only at its lower end and being $\lambda/4$ long, presents a very high impedance at its upper edge. Thus the flow of charges from the inside of the sheath as they emerge from the feeder is reflected from this upper edge as from an open circuit. The standing waves set up on the outside of the sheath over it are images of the charges flowing in the projection of the inner conductor, which is also $\lambda/4$ long. In the *J* form, the same thing takes place except that the stopper section is simply a parallel conductor forming a $\lambda/4$ transformer with the lower part of the projected center conductor which transforms the end-feed impedance back to the coaxial feeder with a value equal to that at the middle of the $\lambda/2$ radiating portion. These three forms of coaxial antennas are specially adapted to vertical antennas which must be mounted at the top of a high flagpole type of support. In practice, the current on the feeder is not completely suppressed, and this point should be considered in the design.

Figure 46*I*, composed of two straight dipoles in space quadrature and excited in phase quadrature, is known as the *turnstile*.[3] When adjusted for equal power and

[1] See reference 302.
[2] See references 258 and 275.
[3] See reference 257.

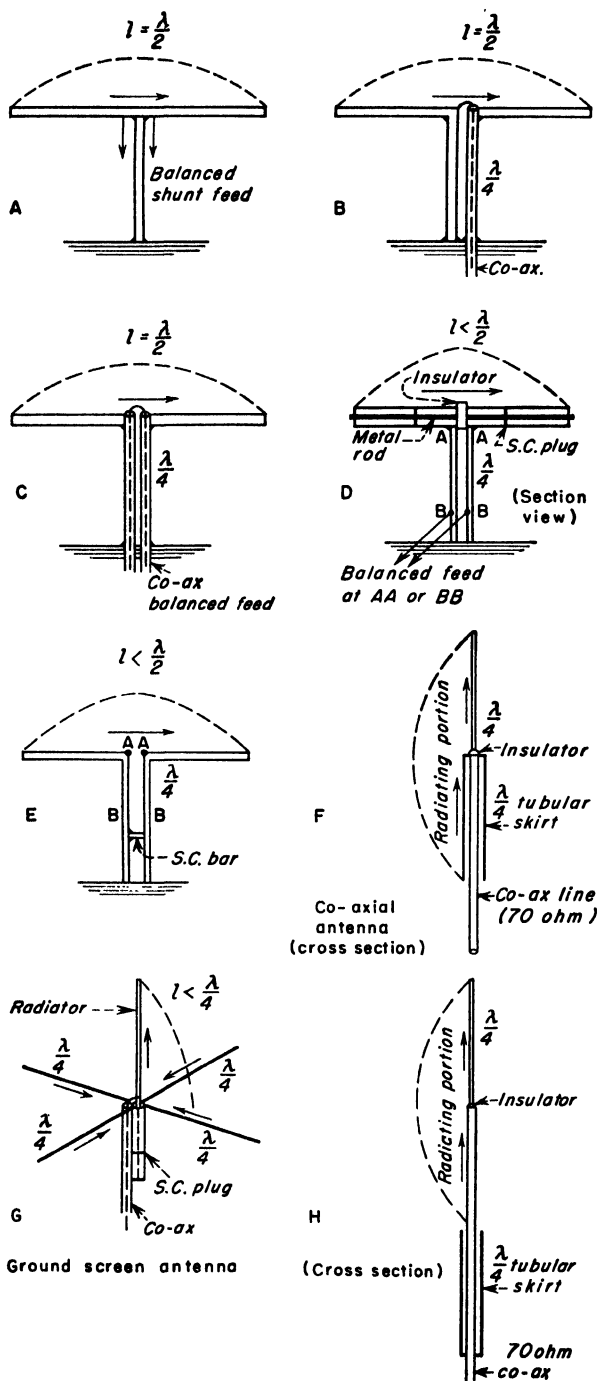



FIG. 46. Basic dipole antennas.

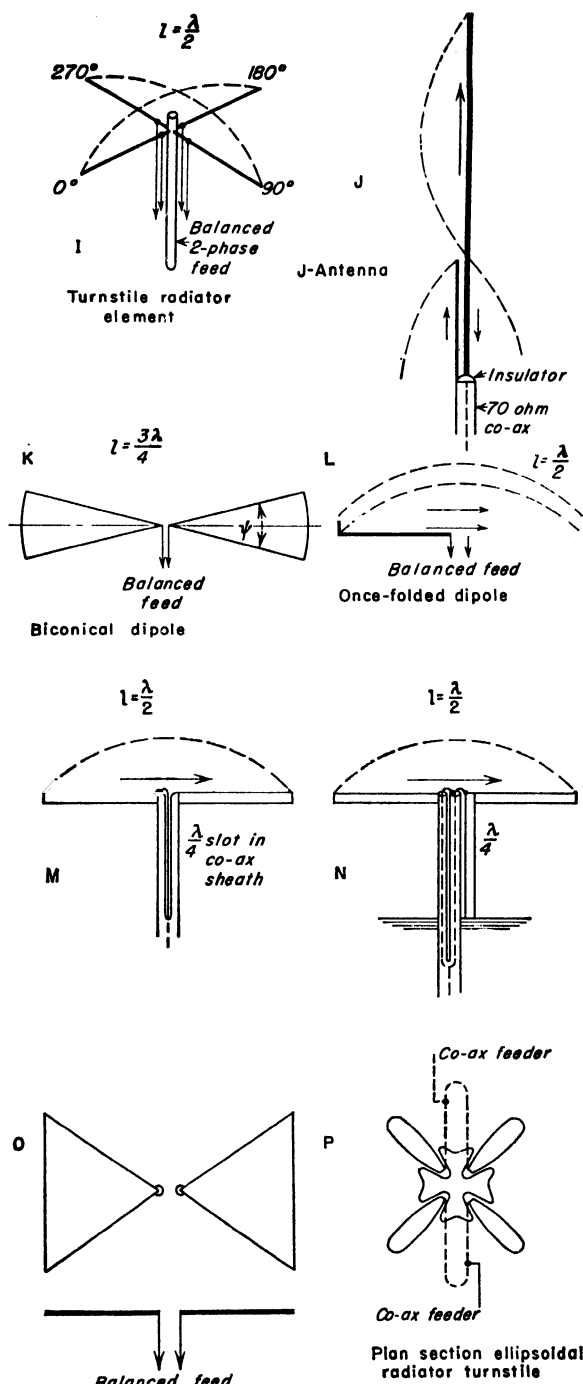



FIG. 46. Basic dipole antennas. (*Continued.*)

exactly 90-deg phase difference between the two quadrature dipoles, the radiation pattern is virtually circular in the plane of the dipoles. As a single element, considerable energy is radiated in the direction normal to the plane of the radiators, and for the specified excitation, the field in this direction is circularly polarized. The polarization can be changed by adjusting either the power divisions or the relative phasing or both. In this case, the radiation pattern in the plane of the turnstile is elliptically shaped. When two or more such radiators are stacked at $\lambda/2$ intervals and successively cophased, maximum radiation occurs in the plane of the radiators. When successive layers are antiphased, directivity normal to the planes of the turnstiles is obtained.[1]

A turnstile element has a certain advantage in providing a greater band width than that of the individual dipoles. This comes about from the method of feeding and phase-shifting whereby the feeder to each radiator is terminated and the two sections excited from a common generator by means of parallel lines from a junction point with the main feeder, one branch being $\lambda/4$ longer than the other.[2] This extra $\lambda/4$ section of line, in addition to providing the 90-deg phase lag, acts as an impedance-inverting section. Both radiators change impedance equally as the frequency is changed from mid-band, but the impedance inversion in one feeder causes a compensating effect at the main junction with the other feeder. This is a valuable property per se.

Figure 46*K* represents the *biconical antenna* whose principal special property is intrinsically greater band-width response than ordinary radiators have.[3]

Figure 46*L* presents again the folded dipole. As an impedance transformer and for its greater cross section providing broader band response than a simple straight dipole, it has special value in certain v-h-f and u-h-f applications.

Figure 46*M* is a variation in principle of Fig. 46*B*, where the central supporting member is a tube forming the outer conductor of a coaxial transmission line. The tube is slotted to a depth of $\lambda/4$, and the inner conductor of the transmission line attaches to one side of the $\lambda/2$ dipole welded to the tube. In general, this type of slotted-coaxial section is both a balance-to-unbalance transformer and an impedance transformer in the ratio of 4:1. That is, the impedance at the balanced terminals (the open end of the slotted section) is four times the impedance of the unbalanced line (assuming it is terminated in a matched load) in shunt with the reactance of the balanced stub formed by the slot in the coaxial line. If the slot is effectively $\lambda/4$ long, its impedance is infinite and the impedance transformation is 4:1. In practice, the characteristic impedance of the coaxial line is modified slightly by the presence of the slot. Unless the diameter of the coaxial line and the slot width are very much less than the wavelength, considerable radiation will occur from the slot.

Figure 46*O* is an example of a dipole, capable of many variations, where linear conductors are replaced by plane sheets of selected shapes for obtaining special characteristics.

In all these forms of electric dipoles, band width can be increased by increasing the radiator cross section when the application is associated with wide-band systems. Their radiation patterns are individually quite similar to any simple straight dipole, except for the turnstile element.

59. Magnetic Dipoles. Certain applications, such as navigational aids, require for their success the radiation of a field which is virtually pure horizontal polarization at any orientation from the antenna. Other less critical applications also use horizontal polarization, the purity of which is not essential. To obtain virtually pure horizontal polarization a vertical magnetic dipole can be used. There are several types in existence, and others appear from time to time. What is wanted in a magnetic dipole is a loop of fairly large electrical diameter to have a high radiation resistance, with a uniformly distributed current around it, all elements of which are in time phase. It has generally been necessary to approach this by configurations made up of bent electric dipoles having standing-wave current distributions. The various forms developed

[1] See reference 291.
[2] A bridge-type network for this purpose is given in reference 295.
[3] See references 16, 265, and 88.

exhibit ingenuity in making magnetic dipoles from electric dipoles, with varying degrees of compromise.

Some of the most practical forms are shown in Fig. 47. In this figure, *a* is a square loop made up of two electric dipoles, folded at their centers and near their ends.[1] When they are attached by means of insulators at the ends, the current at the corner is not zero, though it is much less in amplitude than at the other corners, where the two dipoles are center-fed in balanced antiphased relationship. It is fed by a balanced

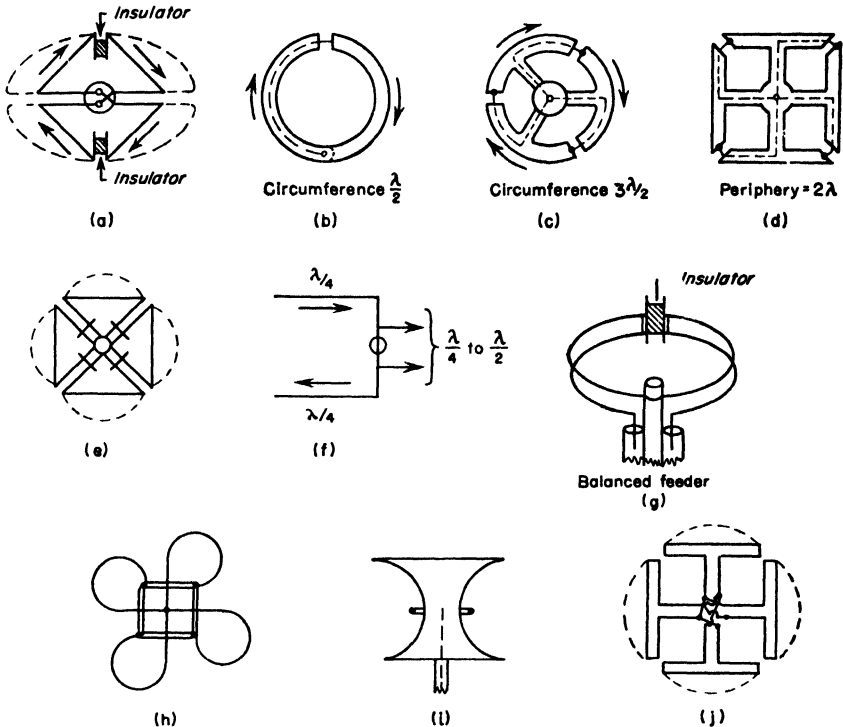



FIG. 47. Magnetic-dipole antennas of various forms.

feeder at the center. This has been used in aviation navigational aids. It is characterized by a nonuniform current distribution which produces an oval pattern, and by relatively low radiation resistance because of its small electrical diameter (approximately $3\lambda/16$ per side). The dipoles are made of wide bars of metal, and the feeder lines are ordinarily shielded up to the two corners where they are energized.

Figures 47*b*, *c*, and *d* are three versions of magnetic dipoles.[2] A circular or polygon loop, sectionalized and fed as shown from a coaxial feeder, permits cophasing the currents in the successive elements for relatively large electrical diameters and provides high values of radiation resistance to be realized. As the loop is increased in diameter, the periphery is sectioned into a larger number of segments, each $\lambda/2$ long, with its own radial feeder. One-half of each radiator acts as the coaxial line for the succeeding radiator.

Figure 47*e* is a square (or alternately a circular) loop radiator $\lambda/2$ per side. The ends are not closed and are connected by radial supporting stubs which are short-

[1] See reference 243.
[2] See reference 285.

circuited $\lambda/4$ from each corner.[1] Thus the current in each side is cophased, while the balanced radial stubs suppress radiation from the antiphased portions of the standing-wave system. This loop can be shunt-excited by a balanced line across one of the radial stubs, or at two such symmetrically opposite points in parallel. The pattern in the plane of the loop is squarish, the maximum field being opposite a side, and is about 1.29 times that opposite a corner.

Figure 47*f* is called the *U antenna*. A special characteristic of this radiator is that it approximates a spherical radiation source, quite unlike all other types of electric and magnetic dipoles shown.

Figure 47*g* is a folded $\lambda/2$ dipole bent in the form of a loop, using balanced feed.[2]

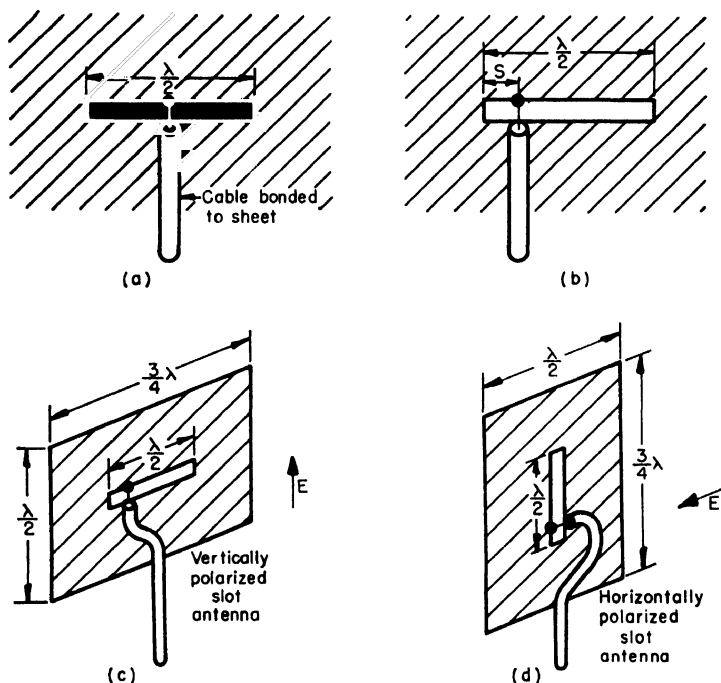



FIG. 48. Slot antennas fed by coaxial transmission lines.

Figure 47*h* is the *cloverleaf* radiator. In this case the tower itself (square) serves as the outer conductors of a concentric line, through the center of which runs the main feeder. Each cloverleaf consists of four loops of rigid tubing, each loop starting at the feeder and terminating on one corner of the tower and bent to include a straight portion and a three-quarters circle. The length of each loop of the cloverleaf is about $\lambda/2$.

Figure 47*i* is a loop radiator located at the focus of a reflector which is a parabola of revolution. The reflector, depending upon its aperture, flattens the radiation pattern. The exciting loop is energized by a coaxial feeder running along the axis of the assembly.

Magnetic dipoles are used as elements in antenna arrays disposed in vertical or horizontal arrangements depending upon the application.

60. Slot Antennas—Current-sheet Radiators. A thin slot cut into a metallic surface with its long dimension approximately $\lambda/2$ is a simple yet efficient radiator by

[1] See reference 321.
[2] See reference 309.

virtue of the currents flowing over the metallic surface.[1] This general type of antenna is called a slot antenna, or current-sheet radiator. It can be energized from a parallel-wire line or from a coaxial line, as shown in Fig. 48. A large mismatch will exist if the latter is used because for $L \doteq \lambda/2$, the input impedance at a pair of terminals at the slot center is approximately 500 ohms (a better match can be achieved by feeding off center).

The general problem of determining the radiation from a slot in a metallic surface can be formulated with the aid of the equivalence theorem.[2] That is, the radiated field may be determined from a distribution of magnetic currents M and electric currents J flowing on a closed surface surrounding the metallic ground plane and radiating as if they were in free space (*i.e.*, in the absence of the ground plane). The current densities are related to the tangential components of E and H on the ground plane, including the slot, by

$$\vec{M} = \vec{E} \times \vec{n}$$
$$\vec{J} = \vec{n} \times \vec{H}$$

where $\vec{E}$ = vector representation of the tangential electric field

$\vec{H}$ = vector representation of the tangential magnetic field

$\vec{n}$ = unit vector normal to the ground plane

The radiation pattern of a slot in an infinite metallic ground plane is a doughnut-shaped pattern of revolution independent of ϕ, as shown in Fig. 49. The radiation is polarized in the ϕ direction. Since radiation occurs on both sides of the metallic surface the polarization changes direction (from the positive ϕ direction to the negative ϕ direction) as one passes through the metallic surface. If L is $\lambda/2$ the pattern shape is independent of ϕ and is given by

$$F(\theta) = \frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}$$

In practice, the ground plane is finite and the radiation pattern is not independent of ϕ. The field produced by the currents on the finite metallic surface as well as the field radiated by the slot must be considered. The pattern is the superposition of these fields.

FIG. 49. Radiation pattern of vertical slot in infinite flat sheet.

By means of Babinet's principle, extended by Booker,[3] many of the problems of slot antennas can be reduced to situations involving complementary linear-dipole antennas for which solutions have already been obtained. A slot antenna and its complementary linear-dipole antenna are illustrated in Fig. 50. The complementary linear antenna consists of a perfectly conducting flat strip identical in size and shape to the slot antenna cut into an infinite perfectly conducting ground plane. The radiation characteristics of these

[1] See reference 9.
[2] See references 12 and 13.
[3] See reference 255.

antennas are similar. That is, the shapes of the radiation patterns of the two are identical, but the polarization is such that E and H are interchanged in the two cases. Also, as noted above, the component of E normal to the ground plane is discontinuous.

The input impedances of a slot antenna Z_s and its complementary linear-dipole antenna Z_d are related by

$$Z_s Z_d = \frac{Z_0^2}{4} = 35{,}476 \text{ ohms}$$

where Z_0 = intrinsic impedance of free space = 120π ohms.

If the slot antenna is enclosed with such a box that zero susceptance is shunted across the input terminals, the input impedance is doubled. This would be the equivalent of placing a "tuned" reflector element behind the complementary dipole.

FIG. 50. A slot antenna and its complementary linear (dipole) antenna.

Methods for exciting sheets of currents on large metallic surfaces, as distinguished from linear distributions, using slots or gaps are shown in Fig 51.

Figure 51*a* is a turnstile assembly of fishtail sheets in which is propagated a continuous distribution of in-phase currents flowing normal to the supporting mast. The figure shows the radiators in one plane only. This design is made of tubular members as a skeleton sheet to reduce wind surface; otherwise it functions as if made of continuous sheets. Its fishtail form was empirically developed to obtain optimum current

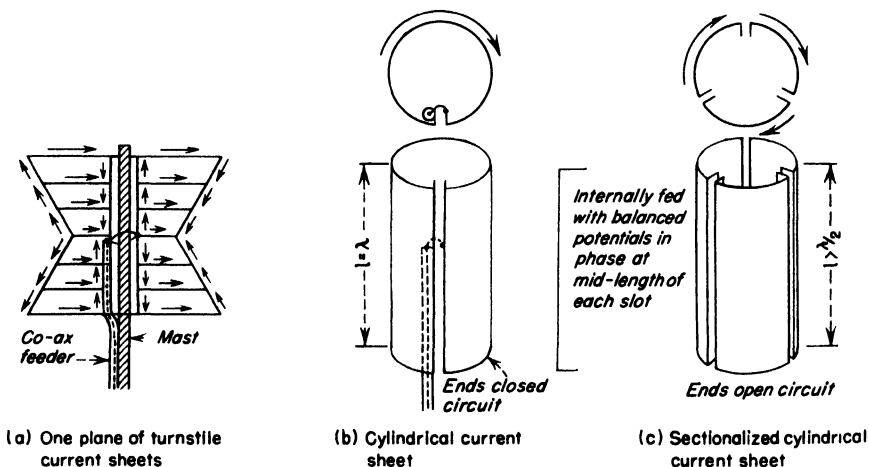



FIG. 51. Current sheet radiators.

distributions and impedance characteristics over a very broad frequency band. Each wing is attached to the mast at top and bottom. A coaxial line rises to the middle of the assembly along one feeder rod to which it is connected. The center conductor emerges at this point and is connected across to the middle of the feeder of the opposite wing. This provides the balanced feed required for the radiators in one plane. Two such planes of radiators in quadrature, excited 90 deg out of phase, comprise a turnstile having a nearly circular pattern normal to the antenna axis.

Figure 51*b* is a current sheet bent into a cylinder (a slotted-cylinder antenna) having a circumference of about $\lambda/2$.[1] The slot acts as a line which is excited with a standing wave by a balanced feeder or by a coaxial feeder arranged for single-end-to-balanced transformation (as in the case preceding). The cylinder can be from one-half (slot

[1] See references 242, 245, 277, 282, 305, 313, 314, 329, 330, and 280.

ends open) to one wavelength long (ends closed) in terms of the phase velocity of a system of this type. The potentials along the feed slot are all cophased, setting up parallel current flow around the circumference of the cylinder at all points. This type of antenna has moderate band width and is very simple to construct.

Figure 51c is a cylindrical (or alternately polygonal) current sheet radiator having three vertical slots. Its theory is similar to the preceding example except for the three

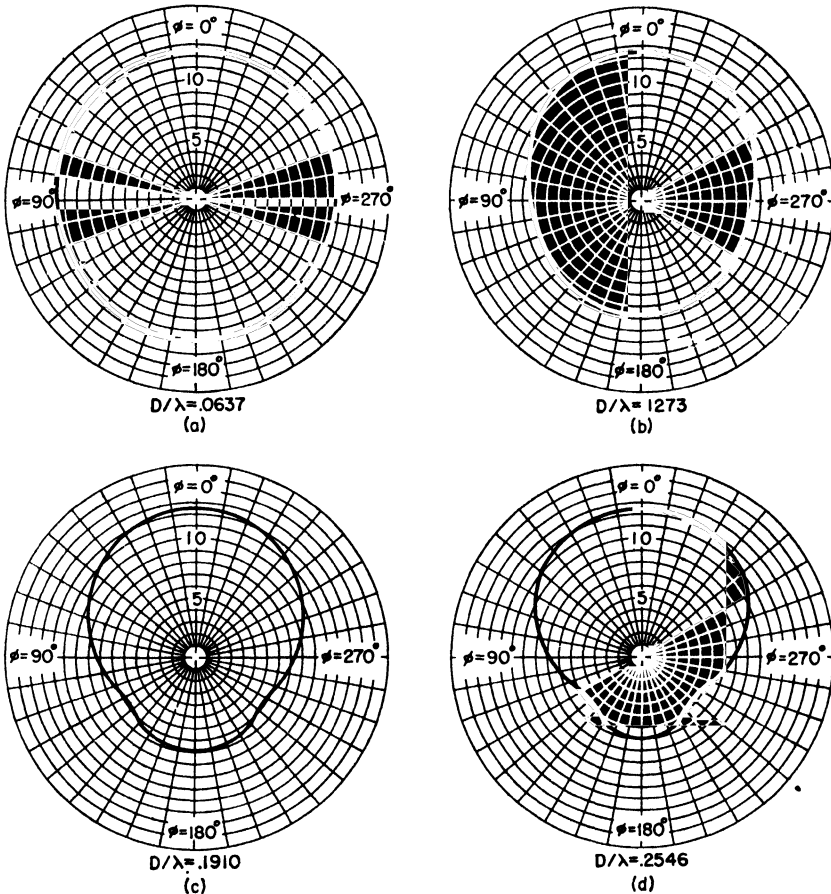



FIG. 52. Patterns for a single slot in a cylinder. The slot is located at $\phi = 0$ deg.

slots, the purpose of which is to obtain a pattern that is a circle. The three slots are fed in parallel at their centers by separate feeders.

Slotted-cylinder antennas have found considerable application for broadcasting a horizontally polarized wave with an omnidirectional pattern in the horizontal plane. They are similar to magnetic dipoles in this respect. Vertical directivity can be obtained by stacking. If the diameter of the cylinder is not too large, a single slot with its long dimension parallel to the axis of the cylinder produces a pattern which, for most purposes, is omnidirectional in the plane normal to the axis. In the orthogonal plane, the pattern is a figure eight. As the diameter increases, the antenna becomes directional. Therefore, the diameter is a simple parameter for pattern control.

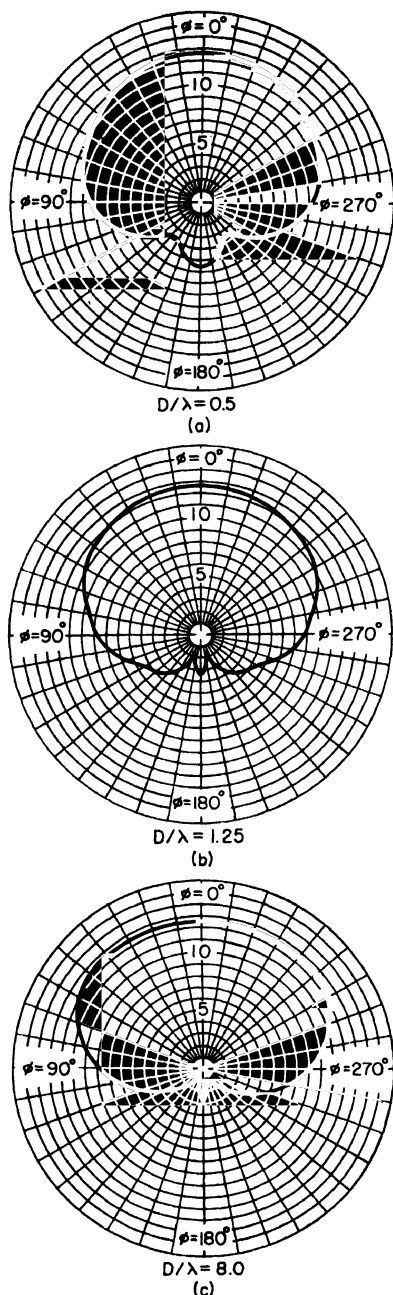



FIG. 53. Patterns of slotted-cylinder antennas having diameters large in terms of the wavelength. Slot is located at $\phi = 0$ deg.

Another method of pattern control consists in arraying a number of slots around the periphery of the cylinder. A wide variety of patterns can be obtained by varying the excitation.

Calculated radiation patterns for a single slot in a cylinder for various cylinder diameters are shown in Figs. 52 and 53.[1] The corresponding variations in relative phase of the radiated field are shown in Fig. 54. Calculated amplitude and phase patterns for two slots diametrically opposite each other, fed equally, and in phase are shown in Figs. 55 and 56 for various cylinder diameters.

61. Elliptical Polarization. The state of elliptical polarization of a plane wave at a point in space may be described in terms of the relative amplitudes and phases of two perpendicular vector field components (electric or magnetic). For example, the ratio of the field components is a complex number whose amplitude is the ratio of the component amplitudes and whose phase is their relative phase difference, and is often used to describe a state of elliptical polarization.

A chart showing the polarization ellipses as a function of the ratio of two orthogonal components (E_2/E_1) and their relative phase difference δ is shown in Fig. 57.[2]

Techniques for handling and representing elliptically polarized waves are covered in a series of excellent articles in the literature.[3]

62. The Helical Antenna.[4] A helix may radiate in many modes. Two of these radiation modes are: (1) the axial-beam mode characteristic of the helical-beam antenna and (2) the normal mode in which the field is maximum in a direction normal to the helix axis.[5] For the latter case, the helix is small compared with a wavelength and is inherently a narrow-band element. In the axial mode the helical-beam antenna has inherent broad-band pattern, impedance, and polarization characteristics. There is a natural adjustment of the phase velocity as a function of frequency so that the fields from each turn add nearly in phase in the end-fire direction over nearly a 2 to 1 frequency range. The terminal impedance is relatively constant over the same frequency range because of the large attenuation of the wave reflected

[1] See reference 313.
[2] See reference 9.
[3] See references 255, 270, 275, 283, 307, and 315.
[4] See references 9 (Chap. 7), 262, and 278.
[5] See reference 324.

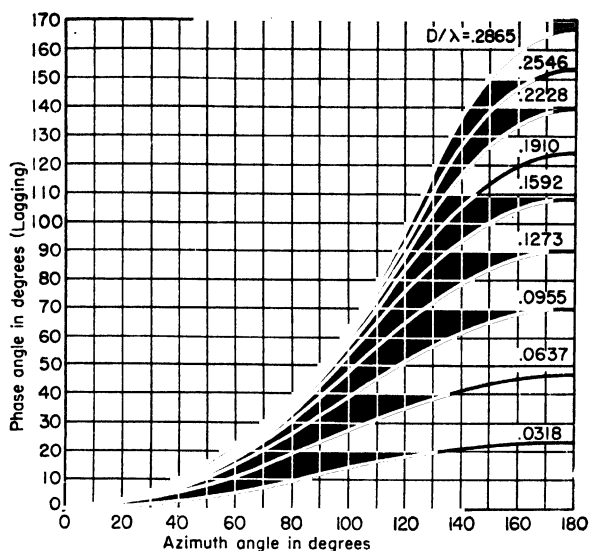



FIG. 54. Curves showing relative phase of the field with azimuth for a single slot in a cylinder. The slot is located at $\phi = 0$ deg in each case.

from the open end of the helix. The polarization is nearly circular over this same range.

Examples of helical-beam antennas are shown in Fig. 58. They are simple to construct. Figures 59 and 60 define the notation used in antenna design. They are

D = helix diameter (center to center)
C = helix circumference $= \pi D$
S = spacing between turns (center to center)
α = pitch angle $= \arctan (S/\pi D)$
L = length of one turn
n = number of turns
A = axial length $= nS$
d = conductor diameter

The dimensions for the helical-beam antenna (for broad-band end-fire operation) are

Pitch angle	$12° < \alpha < 15°$
Circumference	$\frac{3}{4} < C/\lambda_0 < \frac{4}{3}$
Turns	$n > 3$
Upper limit for diameter	$D_u = \dfrac{\sqrt{2S/\lambda_0 + 1}}{\pi}$
Conductor diameter	$0.006 \leqq d/\lambda_0 \leqq 0.05$
Ground-plane size	$>\lambda_0/2$ in diameter

The following antenna characteristics hold only for these dimensions.

The helical-beam antenna has a terminal impedance which is nearly pure resistance with a value between 100 and 200 ohms. To within about ± 20 per cent the terminal impedance of an antenna whose dimensions are within those listed above is given by the empirical relation

$$R = 140C/\lambda_0 \quad \text{ohms}$$

A method of feeding the antenna is shown in Fig. 61. Beyond point Q in this figure, the helix lies in the surface of an imaginary cylinder. Between P and Q, the helix conductor lies in a plane through the helix axis and at approximately the same pitch angle as for the helix proper.

D/λ=.0637 (a)

D/λ=.1273 (b)

D/λ=.1910 (c)

D/λ=.2546 (d)

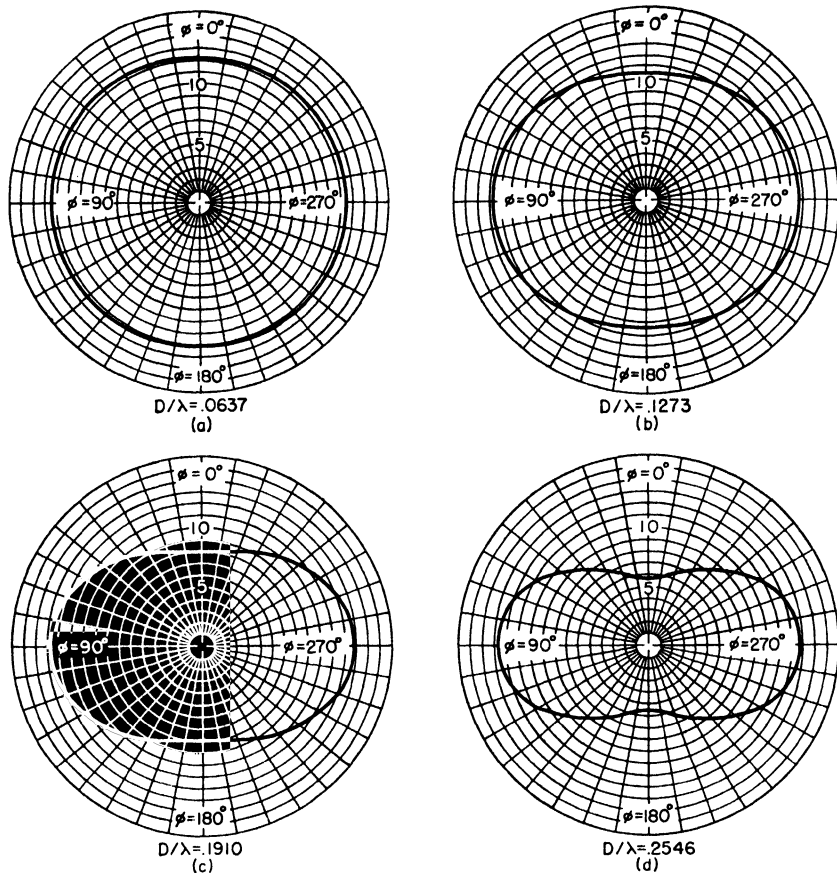



FIG. 55. Patterns for arrays of two diametrically opposed slots, fed equally and in phase. Slots are located at ϕ_1 = 0 deg and ϕ_2 = 180 deg.

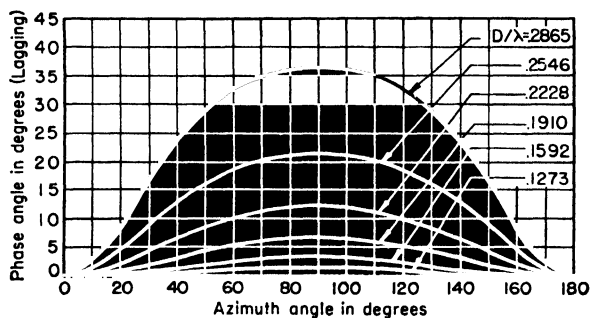



FIG. 56. Variations in relative phase of the fields of arrays of two slots.

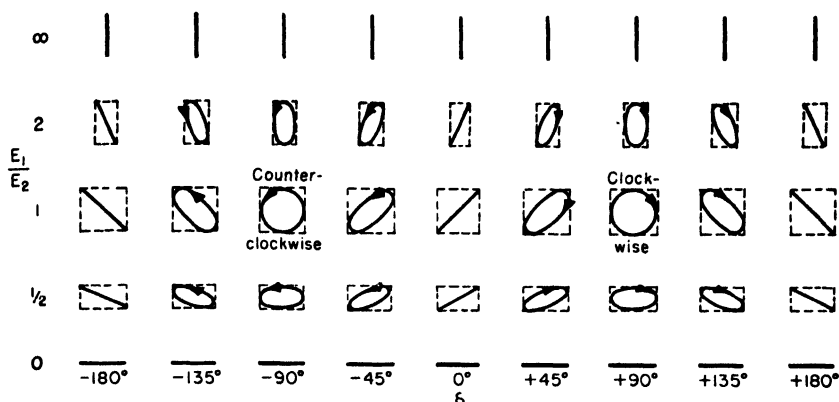



FIG. 57. Polarization ellipses as a function of the ratio E_2/E_1 and phase angle δ (wave approaching).

The relative phase velocity along the helical conductor when the antenna is operating in the beam mode (see conditions above) is

$$\frac{c}{v} = \frac{S/\lambda_0 + [(2n+1)/2n]}{L/\lambda_0}$$

where v = phase velocity along the helical conductor

c = free-space velocity

This corresponds to the Hansen-Woodyard criterion for increased directivity.

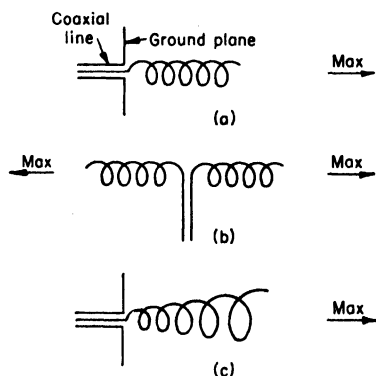



FIG. 58. Arrangement for producing axial mode of radiation.

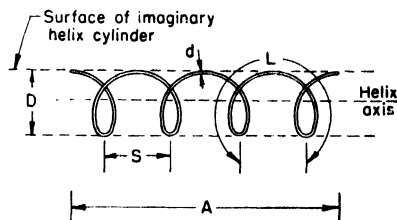



FIG. 59. Helix dimensions.

A good approximation to the far-field pattern can be obtained from an analysis which replaces the helical-beam antenna by an array of n isotropic point sources, each source replacing a single turn of the helix. The normalized pattern function is

$$F(\theta) = \sin\left(\frac{\pi}{2n}\right)\frac{\sin(n\psi/2)}{\sin(\psi/2)}$$

where $\psi = 2\pi[(S/\lambda_0)\cos\theta - (L/\lambda_0)c/v]$

θ = angle measured from helix axis

$\sin(\pi/2n)$ = normalizing factor for increased-directivity condition

Empirical relations for the beam width of a helical-beam antenna (see conditions above) are

$$\text{Half-power BW} \doteq \frac{52}{C/\lambda_0\sqrt{nS/\lambda_0}} \quad \text{deg}$$

$$\text{BW between first nulls} \doteq \frac{115}{C/\lambda_0\sqrt{nS/\lambda_0}} \quad \text{deg}$$

The directivity is given approximately by

$$D \doteq 15(C/\lambda_0)^2 n(S/\lambda_0)$$

The axial ratio (AR) for the helical-beam antenna for the condition of increased directivity is given approximately by

$$AR \doteq \frac{2n + 1}{2n}$$

If n is large, the axial ratio approaches unity and the polarization is nearly circular.

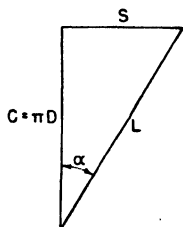



FIG. 60. Relations among cirumference, spacing, turn length, and pitch angle of a helix.

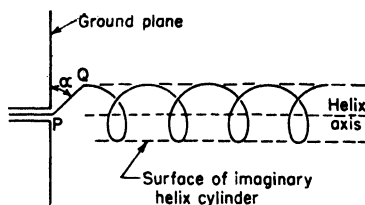



FIG. 61. Terminal arrangement of helical-beam antenna.

63. Antennas for Radiating Simultaneous Vertically and Horizontally Polarized Fields. In certain v-h-f communication applications for aviation, it is required to transmit from the ground station signals of both horizontal and vertical polarization. This can be accomplished in a number of ways with separate antennas energized from a common transmitter but it is preferable to employ a single radiating system.

Figure 62 shows a type of antenna which radiates an omnidirectional circularly polarized field, especially desired in v-h-f aviation communication because of its freedom from orientation effects of the receiving antenna whether the latter be an electric or a magnetic dipole. The four electric dipoles, with equal cophased currents, form a square loop, except that the dipoles are tilted at an angle of 30 deg from their common plane, like sections of a quadruple screw thread.[1]

FIG. 62. Polarization antenna for v.h.f.

64. V-h-f Vertical Arrays. V-h-f f-m and TV broadcasting employs horizontal polarization (in most cases) and requires omnidirectional coverage. To obtain the most effective concentration of radiated power toward the horizon, many forms of vertical arrays have already been evolved using magnetic dipoles, turnstile elements, or current sheets. By increasing the height of the stack, the gain can be adjusted, theoretically, to any desired value. The limitations are set by cost, feeder-line details, and the conditions encountered at the antenna location. The antenna system is

[1] See reference 260.

always located as high above surrounding terrain as possible, using a high tower, building, or natural height. In typical locations the antenna and feeder system is exposed to rigorous weather conditions, and the electrical and mechanical design must be adequate to endure them. In view of the great difficulty of making changes or adjustments after erection, such systems are now usually prefabricated and pretuned.

Gains (with respect to $\lambda/2$ dipole) for a vertical stack of magnetic dipoles fed with equal currents, cophased, and for two forms of turnstile arrays are tabulated below:

Vertical Stack of Magnetic Dipoles

Spacing, deg	Number of magnetic dipoles					
	2	4	6	8	10	12
	Vertical gain, db					
180	1.9	4.5	6.2	7.2	8.0	8.9
360 (optimum)	3.2	6.3	8.4	9.6	10.8	11.7

Vertical Stack of Turnstile Elements

	Number of layers				
	2	4	6	8	10
	Vertical gain, db				
Original turnstile	2.0	3.5	5.0	6.5	8.0
Improved turnstile (double coaxial feed)	1.25	2.75	4.25	5.75	7.2

Several commercial types of vertical arrays for f-m and television broadcasting are in existence, and others are appearing frequently, using various electric and magnetic dipoles as elements. There are arrays of cloverleafs, square loops, circular loops, circular folded dipoles, folded dipole turnstiles, etc.

Where horizontal loops (magnetic dipoles) are arrayed vertically along a common axis and all energized with equal in-phase currents, the normalized vertical pattern factor is

$$F(\theta) = \frac{\sin\theta \sin\left(\frac{n\beta_0 d}{2}\cos\theta\right)}{n \sin\left(\frac{\beta_0 d}{2}\cos\theta\right)}$$

where θ = angle measured from the axis of the array
n = number of loops in the array
$\beta_0 d$ = electrical spacing between loops
= $360\, d/\lambda_0$, deg

This equation does not include the effect of ground reflections, which are usually of minor interest at receiving points along the ground. The figure of merit for an array, including the gain characteristics, depends upon the vertical pattern for the array alone. However, there will be some variation of effective power radiated in azimuth in arrays using elements having noncircular patterns.

Arrays of electric dipoles may also be used, and when directive radiation is required,

they will generally employ elements of this type. Current sheet radiators, singly and in vertical arrays, are also used.

The choice of f-m antennas depends upon many factors, including electrical and mechanical stability, immunity to weather, gain, band width, and, finally, effective radiated power toward the horizon per dollar of installed cost, including tower and feeders.

Television antennas are subject in general to the same requirements as f-m broadcast antennas, with the additional very important one of wide band width, with consequent standing-wave ratio on the feeder of 1.1 or less over the entire video spectrum of the channel.

When a magnetic dipole is made of horizontal electric dipoles, it is interesting to note that a triangle of three electric dipoles produces more nearly circular pattern than any other number.[1]

65. Flat (Planar) Arrays. Radiation control can be effected by planar and by cubical arrays. The cubical array is unusual because it is complicated mechanically and electrically, and almost all desired patterns can be obtained with planar arrays, often with linear arrays. However, the use of reflecting surfaces, in the form of continuous metallic sheets or metallic screens, produces in effect a shallow cubical array because of the radiation from the images in the reflector.

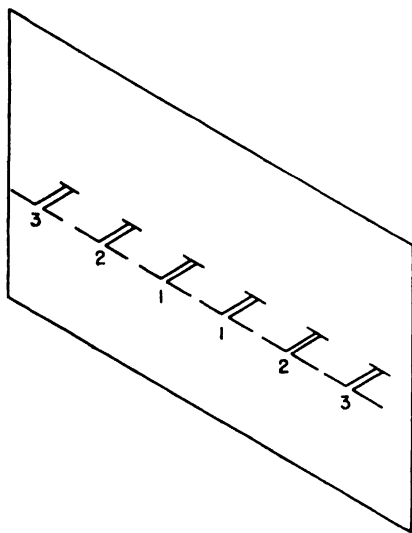



FIG. 63. Colinear array of three pairs of dipoles.

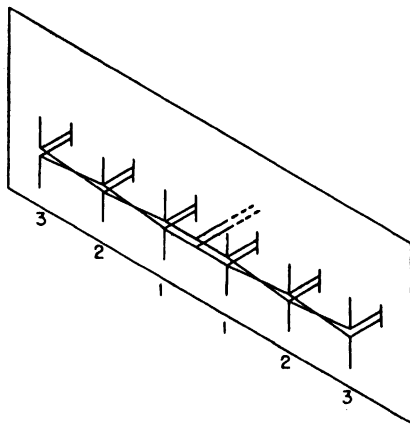



FIG. 64. Linear array of parallel dipoles.

Flat arrays, also called *bedspring* or *billboard arrays*, have been developed in great variety for radar, most of which fall under the following types:

1. Colinear array of dipoles (Fig. 63)
2. Parallel linear array of dipoles (Fig. 64)
3. Two-dimensional (planar) array of dipoles (Fig. 65)

66. Lobe Switching. Lobe switching is a process of producing overlapping patterns alternately and using the equisignal overlap for direction finding (see Fig. 66*a*). To obtain this, the main pattern is a few degrees from the perpendicular to the plane of the array. The methods used to obtain a pattern of the desired shape which is slightly skewed from the perpendicular are notably the following:

1. Dividing the array into two sets of radiators: one producing a pattern which, in terms of azimuth or elevation angle, is an even function, and the other an odd function. By

[1] See reference 266.

inverting the polarity of the feed for the latter, the combined pattern is switched from one side to the other of the normal (Fig. 67).

2. By mechanically shifting the plane of an array periodically in one or two directions.

3. By using two separate arrays oriented to provide the desired overlap using constant excitation if fixed overlapping patterns are wanted and switching the excitation to the two arrays if lobe switching is wanted.

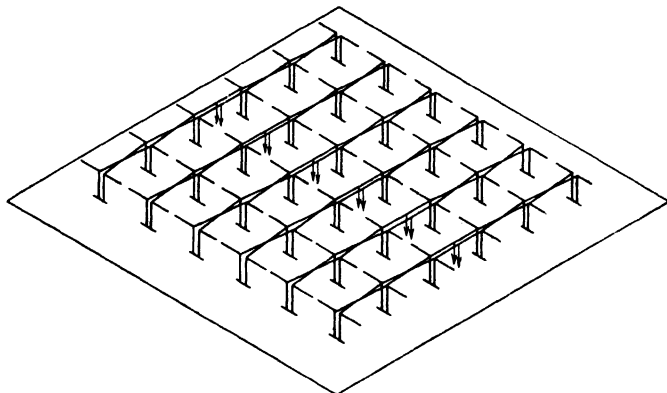



FIG. 65. Square planar array of 36 dipoles.

Arrays have been designed which produce even-function and odd-function patterns simultaneously.[1]

Most of the arrays for communication, navigational radio systems, and radar are composed of combinations of these various characteristics. In a treatise of this scope it is possible only to indicate briefly the principles for this vast field of design.

FIG. 66. Examples of lobe-switching and conical-scanning radiation patterns.

67. Pattern Shaping with Arrays. Pattern shaping is usually accomplished by controlling the relative amplitude and phase of the excitation of the elements of the array (see Sec. 29). If pattern multiplication is applicable, the distributions most commonly used are:

1. *Uniform currents in all radiators.*[2] This distribution produces maximum directivity and relatively high side lobes.

2. *Binomial current distribution symmetrically from center.*[3] This distribution results in zero side-lobe level with a decrease in directivity.

[1] See reference 323.
[2] See references 39 and 93.
[3] See reference 58.

3. *Dolph-Tchebyscheff current distribution (symmetrical).*[1] This distribution optimizes the relation between beam width and side-lobe level. That is, if one is specified the other is minimized.

4. *Fourier current distributions (symmetrical).*[2]

5. *Empirical distributions.*

These distributions are for the control of beam width and the elimination or reduction of secondary lobes.

The simplest arrays use $\lambda/2$ electric dipoles with half spacings of $\lambda/2$, an arrangement closely approximating a continuous sheet distribution of currents. This

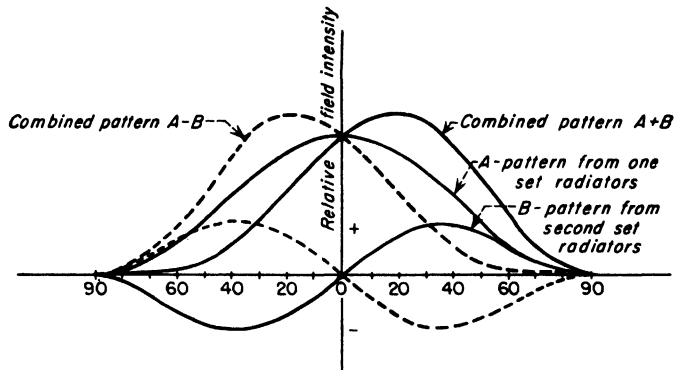



Fig. 67. Simple pattern switching by addition of odd and even pattern functions from separate sets of radiators.

"arrangement" provides convenient impedances and convenient feeder systems. For the same reasons, it is desirable to treat radiators in pairs rather than singly, and the formulas given below are in that form. Other spacings can be used and occasionally are.

In arrays of this sort, the pairs with the greater spacings produce the greatest influence on the sharpness of the main lobe in an equal-current distribution, but they also contribute to the production of large secondary lobes. In radar applications particularly, these secondary lobes are undesirable or positively detrimental, even though the power involved may be small.

Binomial distributions are produced when the radiator currents are graded symmetrically from center as follows:[3]

Array length, λ	No. pairs	Coefficients for pattern formulas					
		A	B	C	D	E	F
		Current ratios from center					
1½	2	0	1.00	0.33			
2	2½	1.00	0.67	0.16			
2½	3	0	1.00	0.50	0.1		
3	3½	1.00	0.75	0.30	0.05		
3½	4	0	1.00	0.60	0.20	0.1	
4	4½	1.00	0.80	0.40	0.11	0.014	
4½	5	0	1.00	0.67	0.29	0.07	0.008

[1] See references 271, 246, 247, 306, and 320.
[2] See reference 326.
[3] See references 58 and 284.

It will be noted from this table that the current amplitudes decrease very rapidly from center. (In odd numbers of radiators, the maximum current is applied to the center radiator; in even numbers, to the central pair.) Outer pairs therefore function almost solely for side-lobe suppression and the central pairs determine the beam width. Actually, radiators with currents less than 10 per cent of maximum can be deleted with small consequences.

68. Fourier Current Distributions. It is frequently desired to produce a particular pattern shape with side lobes suppressed to some specified maximum value where their effect is negligible. To do this, the desired pattern is Fourier-analyzed in terms of spherical harmonics, much as a distorted wave form is analyzed in terms of ordinary harmonics[1] (see also Dolph-Tchebyscheff current distributions). The coefficients from this Fourier analysis determine the current grading by pairs to be used to realize the desired pattern. The number of terms in an infinite series determines the degree of approximation that the designer wishes to approach his ideal objective.

Radiation Patterns for Symmetrical Arrays of Half-wave Dipoles with Half-wave Spacings, Arranged in Colinear, Linear Parallel, and Rectangular Planar Configurations and Located a Distance X from an Infinite Perfectly Conducting Plane Reflector

1. *Equations of Pattern for m Colinear Pairs of Dipoles, Cophased* (Fig. 63):

$$f(\theta) = \underbrace{\left\{\sum_1^m a_m \cos\left[\frac{(2m-1)\pi}{2}\sin\theta\right]\right\}}_{\text{Array factor}} \left\{\underbrace{\left[\frac{\cos\left(\frac{\pi}{2}\sin\theta\right)}{\cos\theta}\right]}_{\text{Dipole factor}} \underbrace{\left[\cos\left(\frac{2\pi x}{\lambda}\right)\cos\theta - \frac{\pi}{2}\right]}_{\text{Reflector factor}}\right\} \qquad (1)$$

$$f(\phi) = \cos\left(\frac{2\pi x}{\lambda}\cos\phi - \frac{\pi}{2}\right) \qquad (2)$$

2. *Equations for Linear Parallel Array of n Pairs of Dipoles, Cophased* (Fig. 64):

$$f(\phi) = \underbrace{\left\{\sum_1^n A_n \cos\left[\frac{(2n-1)\pi}{2}\sin\phi\right]\right\}}_{\text{Array factor}} \underbrace{\left\{\cos\left[\frac{2\pi x}{\lambda}\cos\phi - \frac{\pi}{2}\right]\right\}}_{\text{Reflector factor}} \qquad (3)$$

$$f(\theta) = \underbrace{\left[\frac{\cos\left(\frac{\pi}{2}\sin\theta\right)}{\cos\theta}\right]}_{\text{Dipole factor}} \underbrace{\left[\cos\left(\frac{2\pi x}{\lambda}\cos\theta - \frac{\pi}{2}\right)\right]}_{\text{Reflector factor}} \qquad (4)$$

3. *Planar Array of n Pairs of Rows of m Pairs of Colinear Dipoles, Cophased* (Fig. 65):

$$f(\theta) \equiv \text{Eq. (1)}$$
$$f(\phi) \equiv \text{Eq. (3)}$$

[1] See references 271 and 326.

4. *Lobe-switched Beaver-tail Pattern for Array of Parallel Dipoles, Using n Pairs of Cophased Elements and k Pairs of Interlaced Antiphased Elements* (Fig. 66a):

$$f(\phi) = \left\{ \sum_{1}^{n} A_n \cos\left[\frac{(2n-1)\pi}{2}\sin\phi\right] \right.$$

Even-array factor

$$\left. \pm\, C \sum_{1}^{k} B_k \cos\left[k\pi \sin\phi - \frac{\pi}{2}\right]\right\} \left\{\cos\left[\frac{2\pi x}{\lambda}\cos\phi - \frac{\pi}{2}\right]\right\} \quad (5)$$

Odd-array factor — Reflector factor

$f(\theta) \equiv$ Eq. (4)

5. *Conical Scanning of Ellipsoidal Pattern for Planar Array of n Pairs of Parallel Rows of m Pairs of Colinear Dipoles, Cophased, Interlaced with k Pairs of Parallel Rows of Colinear Dipoles Having l Pairs, Antiphased* (Fig. 66b):

$f(\phi) \equiv$ Eq. (5)

$$f(\theta) = \left\{ \sum_{1}^{m} a_m \cos\left[\frac{(2m-1)\pi}{2}\sin\theta\right] \pm D \sum_{1}^{l} b_l \cos\left[l\pi\sin\theta - \frac{\pi}{2}\right]\right\}$$

Even-array factor — Odd-array factor

$$\left\{\left[\frac{\cos\left(\frac{\pi}{2}\sin\theta\right)}{\cos\theta}\right]\left[\cos\left(\frac{2\pi x}{\lambda}\cos\theta - \frac{\pi}{2}\right)\right]\right\} \quad (6)$$

Dipole factor — Reflector factor

Sequence of switching of antiphased feeders below: C and D are usually equal. The magnitude of the odd-array factor is usually less than that of the even-array factor, by adjustment of the value of C or D.

In these equations, θ and ϕ are measured from the normal to the plane of the array. θ is the angle in the meridian plane (E plane) and ϕ is the equatorial plane of the dipole (H plane).

Interval	C	D
1	+	+
2	−	+
3	−	−
4	+	−
1	+	+

A_n, a_m, B_k, and b_l are Fourier, binomial, uniform, or other systematic coefficients for the current distributions in successive pairs of radiators to obtain desired pattern shapes.

69. Flat Reflecting Screens. A conducting screen, mesh or sheet, of sufficient area, when placed behind a radiating system, acts as a reflecting surface.[1] Use is made of this fact in v-h-f flat arrays. There is diffraction around its edges which sets up fields in the rear of the reflector, and the size used is chosen to keep the stray fields below certain values, depending upon the application. The size is empirical, but in general the higher the directivity of the radiating system, the smaller can be the projecting dimensions of the reflector. In electrically large arrays, the dimensions of the reflector have been as little as $\lambda/2$ more than the radiating system in both length and width. The reflector size depends also on the spacing from the radiators, but this is unimportant for the usual spacings of from 0.1 to 0.25λ. When screens made of wire grids are used, the wires are parallel to the radiators. The spacing between grid wires can be as large as 0.2λ for acceptable forward effectiveness; for better suppression to the rear, smaller spacings are necessary. Designs requiring a high degree of suppression of backward radiation use spacings as small as 0.05λ. The same is true of square mesh

[1] See reference 119.

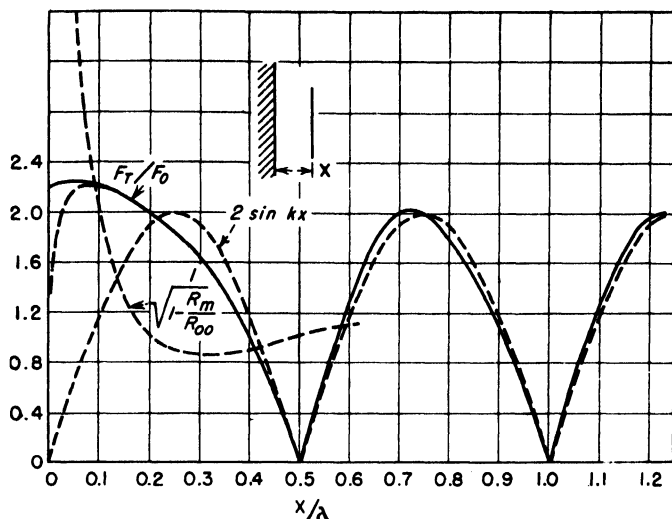



FIG. 68. Transmitting antenna in front of flat sheet. F_t is the field intensity with reflector; F_0 is the field intensity from a free-space dipole; R_m is the mutual impedance between the dipole and its reflector image; R_{00} is the self-resistance of the free-space dipole.

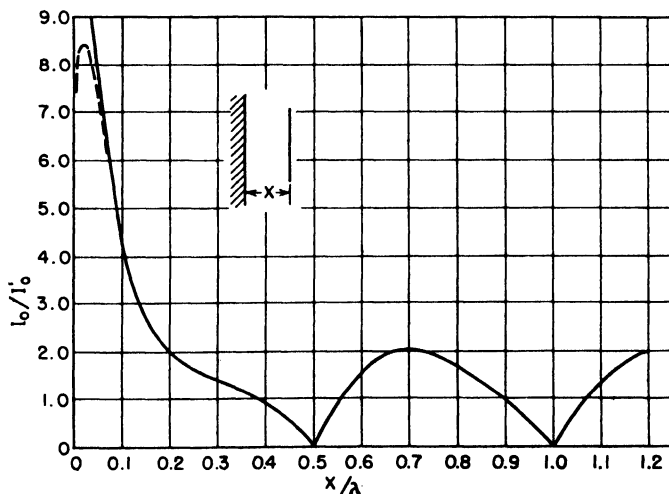



FIG. 69. Receiving antenna in front of flat sheet. I_0' = dipole current without the reflector. I_0 = dipole current with the reflector present.

screens. Screens are preferred to continuous metallic sheets for arrays exposed to winds.

Figure 68 shows the ratio of forward field with a continuous metallic reflector to that without a reflector as a function of spacing for the transmitting case. Figure 69 shows the ratio of received antenna currents with and without a reflector. These approximate the results with screens and grids of finite dimensions.[1]

[1] See reference 119.

70. V-h-f Antenna Measurements. Performance measurements on v-h-f antennas consist principally of (1) measurement of the radiation pattern and (2) measurement of the impedance at the main feeder by means of observing the standing waves on the feeder to be used with the system. For wide-band systems, these are repeated throughout the range of frequencies to be accommodated.

Pattern measurements are made by energizing the radiating system and rotating it on a turntable while observing the field intensity on a recorder at a fixed point. The

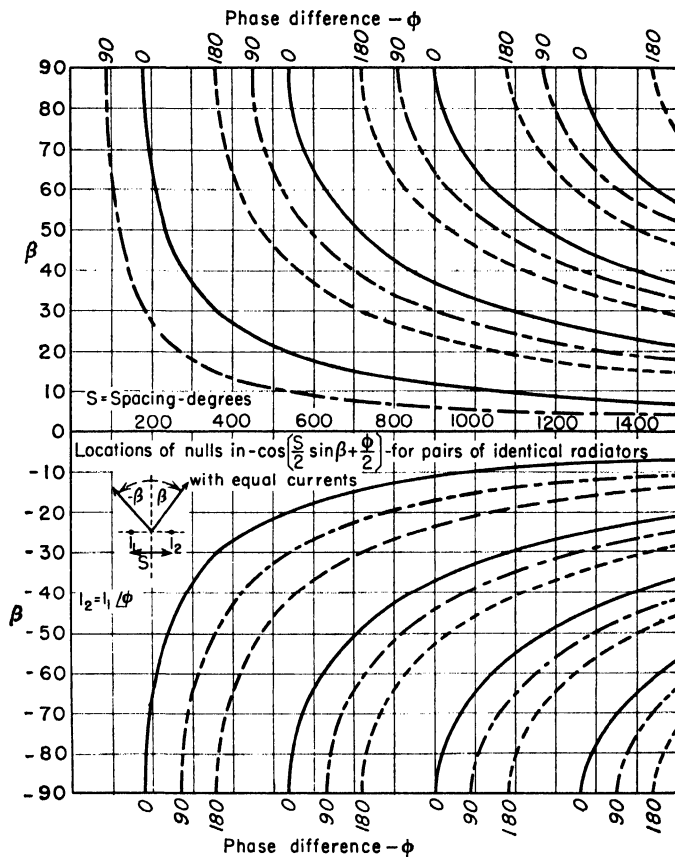



FIG. 70. Curves useful for determining nulls in patterns.

field strength vs. angle is thus examined through 360 deg. Patterns at any orientation can be measured in this way. Errors of large magnitude result when the receiving equipment is moved around a fixed transmitting antenna. The location used should be open and free of objects which will produce reflections in the direction of the receiver, and the receiving antenna should be highly directive toward the turntable and discriminate as much as possible against pickup from the sides. The amplitude response of the detector used must be known so that corrections can be computed for its curvature influence on the observed values. A linear detector indicates relative field intensities directly, and a square-law detector indicates relative power directly. Most detectors fall between these limits, and their empirical results must be corrected.

Incorrect patterns indicate incorrect amplitudes and phases of the currents in the radiators. Improper impedance matches in the feeders must then be located by analysis and corrected. Very little can be done by direct measurement at these feeders, and the final results must be achieved by measurement of the pattern.

The input impedance to the array can be measured by the slotted-line technique or other suitable impedance-measuring instrument.

The band width of an antenna from an input-impedance standpoint is normally determined in terms of some specified standing-wave ratio in the main feeder, assuming virtually perfect match at mid-band. The maximum permitted at the edges of the band depends upon the intended service. For some noncritical applications this can be a standing-wave ratio of 2, while for others it must be as low as 1.1. The standing-wave ratio should be specified when specifying band width.

71. Curves Showing Locations of Nulls in Patterns Such as Fig. 20. In designing directive antennas of various types, use is often made of polar patterns for pairs of radiators, and more extensive arrays are synthesized from such pairs. Figure 70 will be useful in such exploratory work because it shows the angles at which the nulls occur. Interpolation can be made easily between the three phase lines, permitting close estimation of the null positions for any value of phase difference between radiator currents.

72. Values for Certain Basic Patterns for Pairs of Radiators and for the Pattern of the $\lambda/2$ Dipole. Various patterns for extended arrays are synthesized by addition and multiplication of functions for basic elements, among them the $\lambda/2$ dipole pattern and the patterns for certain pairs of radiators. Since $\lambda/2$ spacing between elements of a pair is most frequent in practice, owing to ease of feeding an array, there are tabulated below such combinations as are usually required for cophased and antiphased pairs with integral multiples of the $\lambda/2$ spacing and for the action of neutral reflectors. With these tables, many of the broadside beam antenna patterns can be quickly calculated, including those using various systems of current grading between successive pairs, for both odd and even functions.

S-H-F (MICROWAVE) ANTENNAS (3 TO 30 KMC)

The microwave region is one of transition from the region in which the wavelength is large compared with the dimensions of the components of the system to the optical region in which the wavelengths are very small.[1] Here the wavelengths are of the order of magnitude of the dimensions of easily handled mechanical devices. This leads to a radical modification of earlier antenna techniques and permits the use of complex antenna structures. The techniques of geometrical optics, as well as the theories used at lower frequencies, can be applied to the design of antennas in this region.

73. Horn Antennas.[2] Horn antennas are expanding wave guides of finite length; a relatively large aperture can be obtained. A number of modes are excited in the throat of the horn because of the junction of the wave guide and flare section. However, if the flare angle is not too large, the higher modes are attenuated to a negligible level before free propagation in the horn space is possible.

Many types of antennas can be classified as horn antennas. A few of the more important are shown in Fig. 71. The *sectoral* and *pyramidal horns*[3] have rectangular cross sections and are fed from a rectangular wave guide carrying the dominant $TE_{1,0}$ mode. The *conical horn*[4] has a circular cross section and is fed from a circular guide carrying the $TE_{1,1}$ mode. *Biconical horns*[5] are biconical antennas with large cone angles and hence produce omnidirectional radiation in the plane perpendicular to the cone axes. The *discone antenna*[6] can also be classed as a type of biconical horn.

[1] See reference 15.
[2] See references 9, 12, 13, 15, and 17.
[3] See references 250, 251, 253, 267, and 268.
[4] See references 318 and 355.
[5] See reference 252.
[6] See reference 285.

Table of Values of Functions Corresponding to Radiation Patterns for the Half-wave Dipole, Several Radiator Pairs, and Groups of Pairs

$f(\phi)$	$\frac{\cos\left(\frac{\pi}{2}\sin\phi\right)}{\cos\phi}$	Cophased pairs of circular radiation sources								$\left[\cos\left(\frac{\pi}{2}\sin\phi\right)+\cos\left(\frac{3\pi}{2}\sin\phi\right)\right]0.5$	$\left[\cos\left(\frac{\pi}{2}\sin\phi\right)+\cos\left(\frac{3\pi}{2}\sin\phi\right)+\cos\left(\frac{5\pi}{2}\sin\phi\right)\right]0.38$	$\frac{\cos(90\cos\phi)}{\sin\phi}$	Antiphased pairs of circular radiation sources					
ϕ	$\frac{\lambda}{2}$ dipole	$\cos\left(\frac{\pi}{2}\sin\phi\right)$	$\cos(\Pi\sin\phi)$	$\cos\left(\frac{3\pi}{2}\sin\phi\right)$	$\cos(2\Pi\sin\phi)$	$\cos\left(\frac{5\pi}{2}\sin\phi\right)$	$\cos(3\Pi\sin\phi)$	$\cos\left(\frac{7\pi}{2}\sin\phi\right)$	$\cos(4\Pi\sin\phi)$	2 pairs	3 pairs	$\frac{\lambda}{2}$ dipole	$\cos(36\sin\phi+90)$	$\cos(90\sin\phi+90)$	$\cos(180\sin\phi+90)$	$\cos(360\sin\phi+90)$	$\cos(540\sin\phi+90)$	$\cos(720\sin\phi+90)$
0°	1.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	0	0	0	0	0	0
2.5°	0.997	1.00	0.99	0.97	0.96	0.95	0.93	0.90	0.86	......	......	0.036	−0.04	−0.06	−0.14	−0.26	−0.40	−0.47
5°	0.990	0.99	0.96	0.92	0.85	0.77	0.68	0.56	0.43	0.96	0.89	0.070	−0.08	−0.13	−0.27	−0.52	−0.73	−0.89
7.5°	0.983	0.98	0.92	0.88	0.70	0.49	0.33	0.12	−0.07	......	......	0.103	−0.13	−0.21	−0.40	−0.76	−0.95	−1.00
10°	0.976	0.96	0.86	0.69	0.46	0.20	−0.07	−0.34	−0.57	0.87	0.62	0.138	−0.19	−0.27	−0.515	−0.86	−1.00	−0.81
12.5°	0.962	0.94	0.79	0.53	0.20	−0.11	−0.40	−0.75	−0.90	......	......	0.171	−0.23	−0.33	−0.62	−0.97	−0.93	−0.39
15°	0.950	0.92	0.69	0.35	−0.05	−0.43	−0.76	−0.95	−1.00	0.63	0.28	0.204	−0.27	−0.40	−0.719	−1.00	−0.64	0.11
17.5°	0.933	0.89	0.58	0.17	−0.29	−0.68	−0.94	−0.99	−0.83	......	......	0.240	−0.32	−0.46	−0.80	−0.97	−0.33	0.53
20°	0.916	0.86	0.47	0	−0.55	−0.90	−1.00	−0.81	−0.41	0.43	−0.01	0.271	−0.36	−0.52	−0.88	−0.84	0	0.92
22.5°	0.891	0.83	0.37	−0.18	−0.71	−0.99	−0.94	−0.47	0.10	......	......	0.309	−0.40	−0.57	−0.93	−0.72	0.40	1.00
25°	0.869	0.79	0.27	−0.36	−0.85	−0.99	−0.84	−0.07	0.56	0.21	−0.18	0.345	−0.44	−0.62	−0.97	−0.48	0.74	0.83
27.5°	0.843	0.75	0.14	−0.54	−0.94	−0.88	−0.37	0.32	0.89	......	......	0.380	−0.49	−0.67	−0.99	−0.25	0.95	0.45
30°	0.816	0.71	0	−0.70	−1.00	−0.71	0	0.71	1.00	0	−0.24	0.414	−0.53	−0.71	−1.00	0	1.00	0
32.5°	0.785	0.67	−0.11	−0.83	−0.98	−0.45	0.30	0.93	0.90	......	......	0.453	−0.57	−0.75	−1.00	0.24	0.94	−0.42
35°	0.756	0.62	−0.23	−0.91	−0.90	−0.21	0.63	1.00	0.62	−0.15	−0.17	0.488	−0.60	−0.79	−0.98	0.44	0.78	−0.79
37.5°	0.725	0.57	−0.33	−0.96	−0.78	0.06	0.85	0.92	0.15	......	......	0.523	−0.63	−0.82	−0.95	0.65	0.47	−0.98
40°	0.695	0.53	−0.44	−0.99	−0.63	0.32	0.97	0.71	−0.31	−0.23	−0.05	0.559	−0.66	−0.85	−0.92	0.79	0.17	−0.98
42.5°	0.660	0.48	−0.52	−1.00	−0.45	0.53	1.00	0.40	−0.65	......	......	0.594	−0.69	−0.88	−0.87	0.89	−0.11	−0.82
45°	0.628	0.44	−0.60	−0.99	−0.27	0.74	0.93	0.09	−0.86	−0.27	0.06	0.628	−0.72	−0.89	−0.81	0.96	−0.37	−0.53
47.5°	0.594	0.40	−0.68	−0.95	−0.09	0.88	0.79	−0.21	−0.97	......	......	0.660	−0.76	−0.92	−0.74	1.00	−0.61	−0.15
50°	0.559	0.36	−0.74	−0.90	0.08	0.96	0.60	−0.53	−0.99	−0.27	0.13	0.695	−0.80	−0.93	−0.67	1.00	−0.81	0.21

$\cos(X\sin\phi+90)$ for $-\phi$ angles all signs are reverse of those shown

Table of Values of Functions Corresponding to Radiation Patterns for the Half-wave Dipole, Several Radiator Pairs, and Groups of Pairs. *(Continued)*

$f(\phi)$	$\frac{\cos\left(\frac{\pi}{2}\sin\phi\right)}{\cos\phi}$	Cophased pairs of circular radiation sources								$\left[\cos\left(\frac{\pi}{2}\sin\phi\right)+\cos\left(\frac{3\pi}{2}\sin\phi\right)\right]0.5$	$\left[\cos\left(\frac{\pi}{2}\sin\phi\right)+\cos\left(\frac{3\pi}{2}\sin\phi\right)+\cos\left(\frac{5\pi}{2}\sin\phi\right)\right]0.38$	$\frac{\cos(90\cos\phi)}{\sin\phi}$	Antiphased pairs of circular radiation sources					
ϕ	$\frac{\lambda}{2}$ dipole	$\cos\left(\frac{\pi}{2}\sin\phi\right)$	$\cos(\Pi\sin\phi)$	$\cos\left(\frac{3\pi}{2}\sin\phi\right)$	$\cos(2\Pi\sin\phi)$	$\cos\left(\frac{5\pi}{2}\sin\phi\right)$	$\cos(3\Pi\sin\phi)$	$\cos\left(\frac{7\pi}{2}\sin\phi\right)$	$\cos(4\Pi\sin\phi)$	2 pairs	3 pairs	$\frac{\lambda}{2}$ dipole	$\cos(36\sin\phi+90)$	$\cos(90\sin\phi+90)$	$\cos(180\sin\phi+90)$	$\cos(360\sin\phi+90)$	$\cos(540\sin\phi+90)$	$\cos(720\sin\phi+90)$
52 5°	0 523	0 32	−0 79	−0 83	0 25	0 99	0 38	−0 75	−0 87			0 725	−0 83	−0 95	−0 59	0 96	−0 93	0 51
55°	0 488	0 28	−0 84	−0 76	0 41	0 99	0 14	−0 91	−0 66	−0 24	0 17	0 756	−0 85	−0 96	−0 530	0 90	−0 99	0 77
57 5°	0 453	0 24	−0 88	−0 67	0 56	0 94	−0 07	−0 98	−0 38			0 785	−0 87	−0 97	−0 46	0 83	−0 99	0 91
60°	0 414	0 21	−0 91	−0 59	0 67	0 87	−0 29	−1 00	−0 11	−0 19	0 16	0 816	−0 88	−0 98	−0 40	0 75	−0 95	1 00
62 5°	0 380	0 18	−0 93	−0 51	0 76	0 78	−0 46	−0 96	0 13			0 843	−0 90		−0 34	0 67	−0 88	0 99
65°	0 345	0 15	−0 95	−0 43	0 83	0 68	−0 63	−0 86	0 37	−0 14	0 13	0 869	−0 92	−0 99	−0 29	0 574	−0 78	0 94
67 5°	0 309	0 12	−0 97	−0 35	0 88	0 56	−0 75	−0 74	0 57			0 891	−0 94		−0 24	0 47	−0 67	0 84
70°	0 271	0 10	−0 98	−0 28	0 92	0 45	−0 84	−0 62	0 73	−0 09	0 09	0 916	−0 95	−1 00	−0 19	0 38	−0 55	0 69
72 5°	0 240	0 07	−0 98	−0 22	0 96	0 36	−0 90	−0 49	0 85			0 933	−0 96	−1 00	−0 15	0 29	−0 44	0 53
75°	0 204	0 06	−0 99	−0 16	0 98	0 28	−0 94	−0 37	0 91	−0 05	0 06	0 950	−0 97	−1 00	−0 11	0 22	−0 34	0 39
77 5°	0 171	0 04	−0 99	−0 10	0 99	0 20	−0 97	−0 25	0 96			0 962		−1 00	−0 08	0 16	−0 23	0 28
80°	0 138	0 03	−1 00	−0 05	1 00	0 14	−0 99	−0 15	0 98	−0 01	0 04	0 976	−0 98	−1 00	−0 05	0 10	−0 14	0 17
82 5°	0 103	0 02	−1 00	−0 03	1 00	0 09	−1 00	−0 09	0 99			0 983		−1 00	−0 03	0 06	−0 08	0 11
85°	0 070	0 01	−1 00	−0 01	1 00	0 05	−1 00	−0 04	1 00	0	0 02	0 9[illegible]	−0 99	−1 00	−0 02	0 03	−0 04	0 07
87 5°	0 036	0	−1 00	0	1 00	0 02	−1 00	−0 01	1 00			0 999		−1 00	−0 01	0 01	−0 01	0 03
90°	0	0	−1 00	0	1 00	0	−1 00	0	1 00	0	0	1 000	−1 00	−1 00	0	0	0	0

cos (X sin φ + 90) for −φ angles all signs are reverse of those shown

The sectoral horn is flared in one dimension only, either in the plane parallel to the electric vector (E plane) or in the plane parallel to the magnetic vector (H plane). Such a horn produces a fan-shaped beam sharp in the plane of the flare and broad in the orthogonal plane.

Horns are often used where antennas with medium directive gains are required. For example, a pyramidal horn with a directivity of 400 (26 db) will be approximately

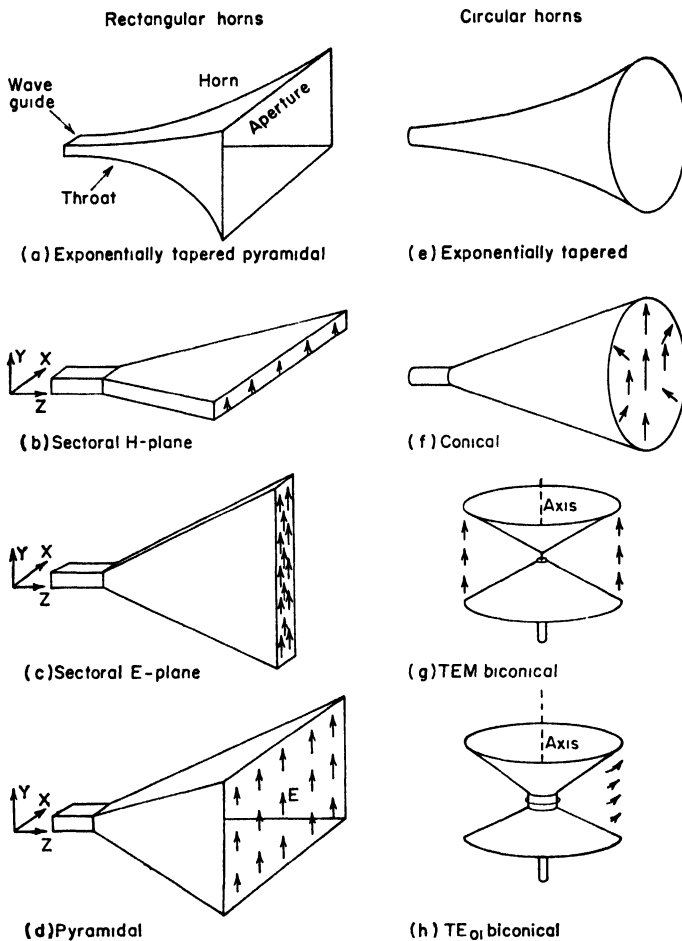



FIG. 71. Types of rectangular and circular horn antennas.

20λ long (which at X band is about 2 ft). Horns are used as highly accurate secondary standards of directive gain because the measured and theoretical values are in very close agreement. They are also employed as the primary feed to illuminate a reflector or lens. In this case the input impedance and efficiency of illumination, rather than directive gain of the horn, are of primary importance.

For small flare angles the field-strength variations across the aperture of a rectangular horn are similar to the variations of the $TE_{1,0}$ mode across the rectangular wave guide feeding the horn. Therefore the aperture field is essentially uniform in magnitude in the E plane and sinusoidally distributed in magnitude in the H plane. This is

the assumption used in pattern calculations, and its validity is borne out in practice by the radiation patterns that result in the E and H planes.

To ensure in-phase fields across the aperture, and therefore high directivity, a very long horn with small flare angle is required; practically, this is inconvenient. An *optimum* horn is defined as one for which the flare angle(s) are chosen so that, for a given length of horn, the directivity is a maximum. For a given horn length, the aperture area, hence directivity, increases with the flare angle. However, beyond a certain flare angle this is overshadowed by phase-error effects in the aperture. If the slant length of the horn is l, the aperture A, and the flare angle θ, as shown in Fig. 72, the difference in path length between the center and edge δ is

$$\delta = l(1 - \cos \theta/2) \doteq A^2/8l$$

FIG. 72. Phase error in the aperture of a horn antenna.

The corresponding phase difference is very nearly $\beta_0\delta$ because the wavelength at the mouth of the horn is essentially equal to the free-space wavelength (β_0 = free-space phase constant). This is the fundamental limitation of all horn antennas.

To a first approximation, δ for optimum horns is independent of the directivity of the horn but is different for the various types of horns. A table of approximate values of δ for optimum horns obtained from theory and measurement is given below.[1]

Type of Horn	δ
H-plane sectoral, or H plane of pyramidal horn	0.40
E-plane sectoral, or E plane of pyramidal horn	0.25
Conical horn	0.32

Note that the tolerance in the H plane is greater than for the E plane. This is true because the contributions from the fields near the sides of the horn to the radiated field are considerably less in the H plane than in the E plane because of the sinusoidal H-plane distribution.

It has been found that, if the aperture in both planes is greater than about 1 to 2λ, the pattern in one plane is essentially independent of the aperture dimensions in the other plane. A set of radiation patterns of pyramidal-horn antennas has been measured as a function of the E-plane flare angle θ, H-plane flare angle ϕ (from 0 to 50 deg), and radial length of the horn R (from 0 to 50λ).[2] These are shown in Figs. 73 and 74. All patterns have been normalized to the same maximum value, hence give no indication of relative directivity. These figures indicate the variety of patterns that can be obtained from a horn antenna and can serve as a guide in horn design. The pattern band width for a particular horn design can be determined from the patterns. Since the radiation patterns in one plane are independent of the aperture dimensions in the other plane these patterns are also applicable to E-plane and H-plane sectoral horns.

Approximate expression for the half-power beam width and the beam width between nulls of the main beam in degrees for optimum horns with relative large apertures are given below.[3] The E- and H-plane beam widths of an optimum pyramidal horn can be found from the formulas for the respective sectoral horn.

Horn	Angle between nulls, deg	Angle between half-power points, deg
Optimum E-plane sectoral horn	$115\lambda/b$	$56\lambda/b$
Optimum H-plane sectoral horn	$172\lambda/a$	$67\lambda/a$
Optimum conical horn:		
E plane	...	60λ/diam
H plane	......	70λ/diam

[1] See references 9 and 355.
[2] See reference 369.
[3] See references 9 (p. 380) and 355.

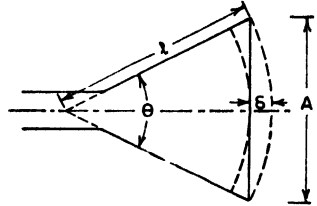
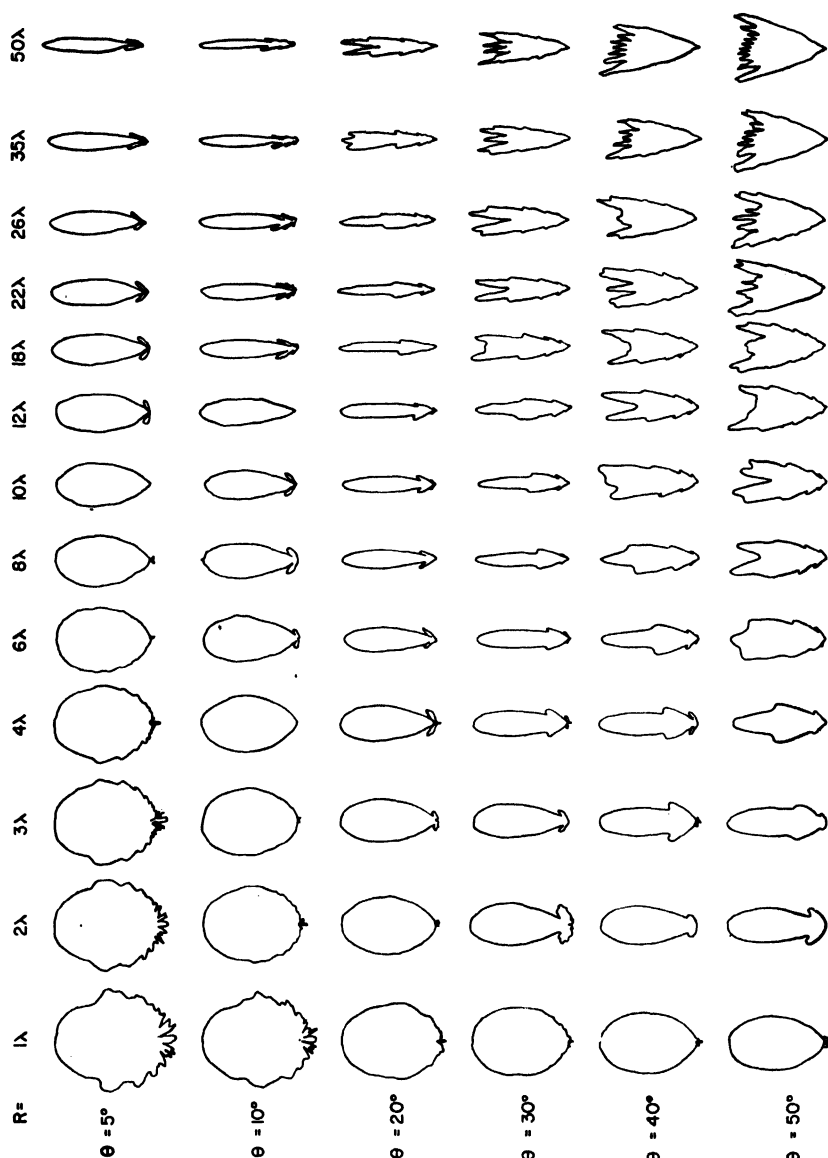



FIG. 73. Electric-plane patterns of a rectangular horn antenna as a function of radial horn length R and E-plane flare angle θ.

The field radiated by a properly constructed rectangular horn is linearly polarized. In the E and H planes the direction of polarization is parallel to the electric field in the aperture. Circular polarization can be obtained if a square horn is fed from a square wave guide propagating orthogonal (space and time) TE modes.

The directivity of a horn antenna can be calculated quite accurately from the physical dimensions of the horn and use of curves or tables.[1] These data are based upon the

[1] See references 12, 330, and 351.

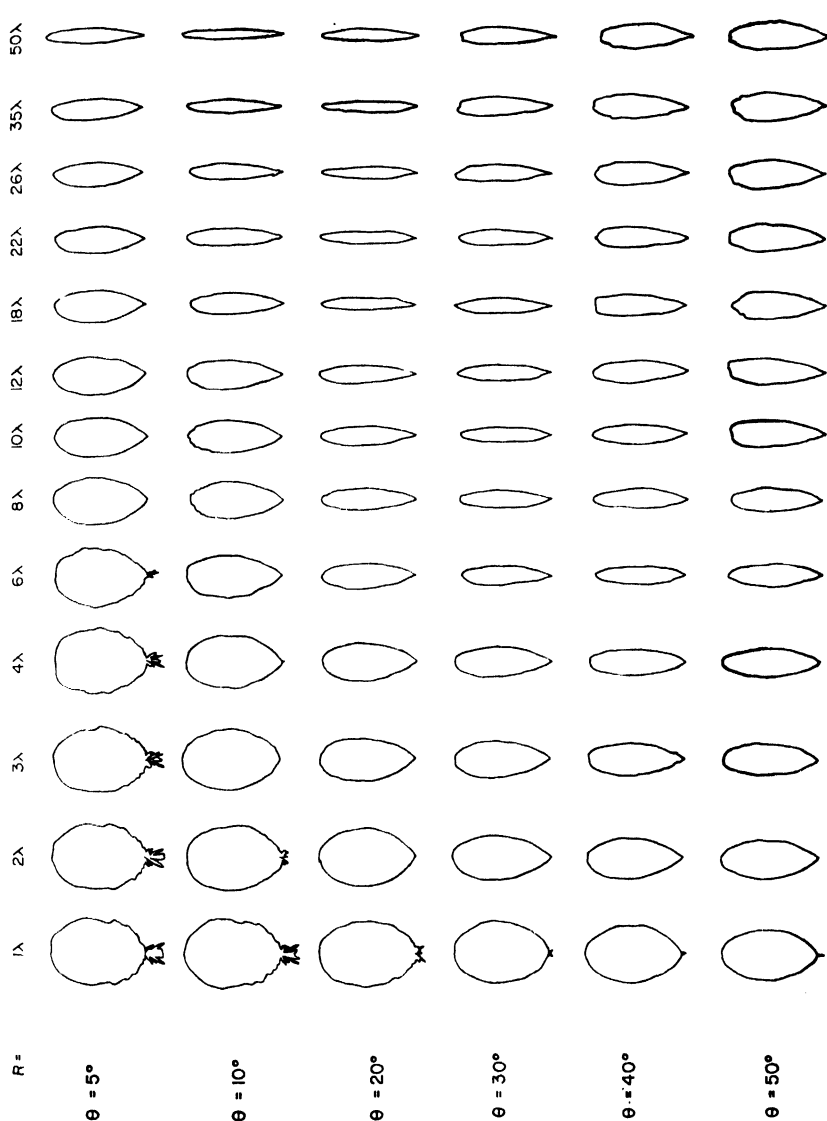



FIG. 74. Magnetic-plane patterns of a rectangular horn antenna as a function of radial horn length R and H-plane flare angle ϕ.

assumption that the field in the horn aperture is the same as though the sides were infinitely long. Experimental measurements have indicated that the error due to this assumption is negligibly small.

The directivity of a horn antenna may be determined with the aid of the tables[1] given below and Figs. 75 to 77. To find the directivity of a pyramidal horn calculate the quantities

$$A = a\sqrt{50/l_H} \qquad \text{and} \qquad B = b\sqrt{50/l_E}$$

[1] See reference 331.

where a = H-plane aperture dimension in free-space wavelengths
b = E-plane aperture dimension in free-space wavelengths
l_H = H-plane slant length in free-space wavelengths
l_E = E-plane slant length in free-space wavelengths
and find the corresponding values for G_H and G_E in the tables for the proper values of A and B, respectively. Then the directivity g is given by

$$g = \frac{G_E G_H}{\frac{32}{\pi}\sqrt{50/l_H}\sqrt{50/l_E}}$$

For an H-plane sectoral horn calculate

$$A = a\sqrt{50/l_H}$$

and find G_H from the tables. Then,

$$g_H = \frac{bG_H}{\sqrt{50/l_H}}$$

For an E-plane sectoral horn calculate

$$B = b\sqrt{50/l_E}$$

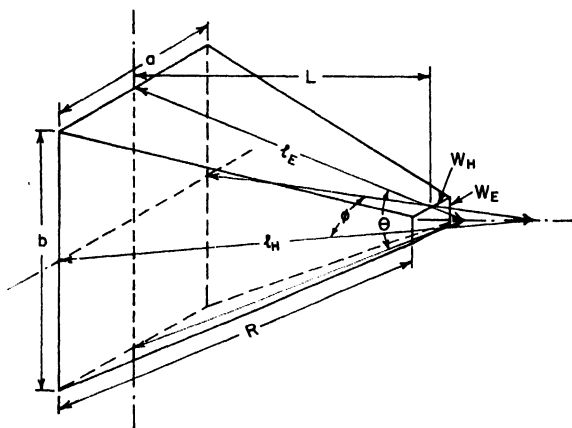



FIG. 75. Dimensions of the pyramidal-horn antenna.

G_H as a Function of H-plane Aperture Dimension A

A	G_H	A	G_H	A	G_H	A	G_H	A	G_H	A	G_H	A	G_H
2.0	20.370	4.6	46.635	7.2	71.291	9.8	90.633	12.4	99.019	15.0	92.591	17.6	75.416
2.1	21.387	4.7	47.628	7.3	72.164	9.9	91.195	12.5	99.052	15.1	92.066	17.7	74.701
2.2	22.402	4.8	48.619	7.4	73.031	10.0	91.740	12.6	99.062	15.2	91.528	17.8	73.991
2.3	23.422	4.9	49.609	7.5	73.889	10.1	92.270	12.7	99.051	15.3	90.972	17.9	73.282
2.4	24.439	5.0	50.595	7.6	74.739	10.2	92.781	12.8	99.012	15.4	90.400	18.0	72.581
2.5	25.452	5.1	51.578	7.7	75.580	10.3	93.274	12.9	98.953	15.5	89.822	18.1	71.886
2.6	26.471	5.2	52.559	7.8	76.413	10.4	93.751	13.0	98.871	15.6	89.214	18.2	71.199
2.7	27.488	5.3	53.536	7.9	77.236	10.5	94.208	13.1	98.763	15.7	88.601	18.3	70.516
2.8	28.501	5.4	54.512	8.0	78.049	10.6	94.646	13.2	98.638	15.8	87.976	18.4	69.847
2.9	29.518	5.5	55.475	8.1	78.854	10.7	95.067	13.3	98.486	15.9	87.337	18.5	69.183
3.0	30.532	5.6	56.449	8.2	79.644	10.8	95.470	13.4	98.309	16.0	86.688	18.6	68.534
3.1	31.545	5.7	57.418	8.3	80.427	10.9	95.848	13.5	98.114	16.1	86.026	18.7	67.891
3.2	32.560	5.8	58.377	8.4	81.196	11.0	96.207	13.6	97.894	16.2	85.355	18.8	67.262
3.3	33.573	5.9	59.334	8.5	81.956	11.1	96.547	13.7	97.654	16.3	84.677	18.9	66.643
3.4	34.579	6.0	60.286	8.6	82.703	11.2	96.869	13.8	97.387	16.4	83.990	19.0	66.038
3.5	35.595	6.1	61.232	8.7	83.440	11.3	97.168	13.9	97.101	16.5	83.319	19.1	65.447
3.6	36.605	6.2	62.176	8.8	84.164	11.4	97.446	14.0	96.793	16.6	82.594	19.2	64.871
3.7	37.612	6.3	63.115	8.9	84.875	11.5	97.702	14.1	96.464	16.7	81.888	19.3	64.305
3.8	38.622	6.4	64.046	9.0	85.567	11.6	97.938	14.2	96.113	16.8	81.179	19.4	63.758
3.9	39.629	6.5	64.975	9.1	86.250	11.7	98.149	14.3	95.740	16.9	80.461	19.5	63.222
4.0	40.633	6.6	65.896	9.2	86.923	11.8	98.342	14.4	95.348	17.0	79.742	19.6	62.703
4.1	41.637	6.7	66.810	9.3	87.579	11.9	98.510	14.5	94.936	17.1	79.023	19.7	62.201
4.2	42.645	6.8	67.720	9.4	88.221	12.0	98.658	14.6	94.504	17.2	78.301	19.8	61.711
4.3	43.639	6.9	68.628	9.5	88.844	12.1	98.783	14.7	94.054	17.3	77.578	19.9	61.243
4.4	44.641	7.0	69.518	9.6	89.460	12.2	98.882	14.8	93.586	17.4	76.854	20.0	60.788
4.5	45.639	7.1	70.407	9.7	90.053	12.3	98.965	14.9	93.095	17.5	76.134		

G_E as a Function of E-plane Aperture Dimension B

B	G_E	B	G_E	B	G_E	B	G_E	B	G_E	B	G_E	B	G_E
2.0	20.362	4.6	46.397	7.2	69.123	9.8	81.301	12.4	73.784	15.0	46.499	17.6	19.910
2.1	21.381	4.7	47.362	7.3	69.847	9.9	81.426	12.5	73.041	15.1	45.268	17.7	19.316
2.2	22.395	4.8	48.326	7.4	70.555	10.0	81.518	12.6	72.265	15.2	44.040	17.8	18.767
2.3	23.410	4.9	49.283	7.5	71.248	10.1	81.581	12.7	71.459	15.3	42.813	17.9	18.264
2.4	24.425	5.0	50.233	7.6	71.923	10.2	81.611	12.8	70.621	15.4	41.593	18.0	17.805
2.5	25.440	5.1	51.181	7.7	72.586	10.3	81.609	12.9	69.753	15.5	40.379	18.1	17.395
2.6	26.456	5.2	52.123	7.8	73.219	10.4	81.575	13.0	68.856	15.6	39.174	18.2	17.030
2.7	27.472	5.3	53.057	7.9	73.841	10.5	81.510	13.1	67.931	15.7	37.982	18.3	16.714
2.8	28.481	5.4	53.985	8.0	74.441	10.6	81.408	13.2	66.980	15.8	36.801	18.4	16.445
2.9	29.490	5.5	54.908	8.1	75.025	10.7	81.277	13.3	66.001	15.9	35.636	18.5	16.223
3.0	30.503	5.6	55.821	8.2	75.585	10.8	81.110	13.4	64.997	16.0	34.488	18.6	16.048
3.1	31.511	5.7	56.728	8.3	76.127	10.9	80.909	13.5	63.969	16.1	33.359	18.7	15.921
3.2	32.518	5.8	57.626	8.4	76.645	11.0	80.676	13.6	62.917	16.2	32.250	18.8	15.839
3.3	33.527	5.9	58.517	8.5	77.142	11.1	80.405	13.7	61.844	16.3	31.164	18.9	15.804
3.4	34.530	6.0	59.401	8.6	77.616	11.2	80.104	13.8	60.748	16.4	30.104	19.0	15.814
3.5	35.534	6.1	60.272	8.7	78.065	11.3	79.765	13.9	59.635	16.5	29.069	19.1	15.870
3.6	36.534	6.2	61.134	8.8	78.492	11.4	79.393	14.0	58.501	16.6	28.063	19.2	15.967
3.7	37.531	6.3	61.987	8.9	78.892	11.5	78.987	14.1	57.351	16.7	27.086	19.3	16.108
3.8	38.530	6.4	62.828	9.0	79.269	11.6	78.545	14.2	56.188	16.8	26.142	19.4	16.289
3.9	39.524	6.5	63.659	9.1	79.619	11.7	78.068	14.3	55.008	16.9	25.232	19.5	16.521
4.0	40.515	6.6	64.477	9.2	79.944	11.8	77.559	14.4	53.816	17.0	24.355	19.6	16.769
4.1	41.504	6.7	65.285	9.3	80.240	11.9	77.014	14.5	52.614	17.1	23.515	19.7	17.064
4.2	42.490	6.8	66.080	9.4	80.510	12.0	76.435	14.6	51.402	17.2	22.713	19.8	17.394
4.3	43.472	6.9	66.862	9.5	80.752	12.1	75.822	14.7	50.183	17.3	21.951	19.9	17.755
4.4	44.450	7.0	67.630	9.6	80.964	12.2	75.176	14.8	48.959	17.4	21.228	20.0	18.147
4.5	45.425	7.1	68.385	9.7	81.146	12.3	74.497	14.9	47.731	17.5	20.548		

and find G_E from the tables. Then,

$$g_E = \frac{aG_E}{\sqrt{50/l_E}}$$

The directivity of an *optimum* H-plane sectoral horn is

$$g_H = 7.86\ ab$$

and the directivity of an *optimum* E-plane sectoral horn is

$$g_E = 8.0\ ab$$

The directivity of a large uniformly illuminated rectangular aperture of dimensions a and b (in wavelengths) is $g_0 = 4\pi ab$. Hence, the directivity of the sectoral horns relative to the uniformly illuminated aperture (sometimes called *aperture efficiency*) is approximately

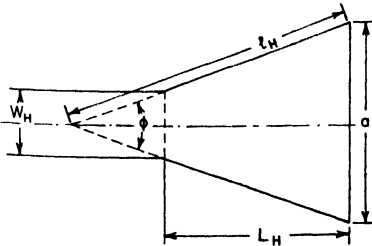



Fig. 76. Dimensions of the H-plane sectoral-horn antenna.

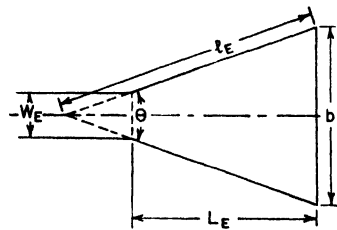



Fig. 77. Dimensions of the E-plane sectoral-horn antenna.

0.62 (62 per cent) for the optimum H-plane horn and 0.64 (64 per cent) for the optimum E-plane horn.

Since horn antennas are nonresonant structures, they present a fairly good termination for wave guides over a broad band of frequencies. The input admittance observed in the wave-guide feed depends upon the admittance of the aperture, the transforming properties of the flared section, and the transforming properties of the junction with the wave guide.[1]

If a horn is to be used in a communications system, the important considerations are usually maximum directivity and minimum size (*i.e.*, an optimum horn), while if a horn is to be used as a primary feed, the efficiency of illumination of the reflector or lens, not directivity, is important. There are other applications where the important design considerations differ from these two. For example, in the field of antenna measurements it is usually required that a horn be designed to illuminate a specified sector at a minimum range with a minimum of variation in amplitude (and phase).

74. Horn Design. A procedure for designing an optimum pyramidal horn whose sides meet the wave guide in a common plane is as follows.[2] If the directivity g is specified, then the dimensions of the horn in the E plane are

$$b = 0.346\sqrt{g}$$
$$l_E = 0.0576g$$
$$\sin\frac{\theta}{2} = \frac{b}{2l_E} = \frac{3.0}{\sqrt{g}}$$

In the H plane the dimensions are

$$a = 0.468\sqrt{g}$$
$$l_H = \frac{a}{a - w_H}\sqrt{\left[l_E^2 - \left(\frac{b}{2}\right)^2\right]\left[\frac{b - w_E}{b}\right]^2 + \left[\frac{a - w_H}{2}\right]^2}$$
$$\sin\frac{\phi}{2} = \frac{a}{2l_H}$$

where a = H-plane aperture dimension
b = E-plane aperture dimension
l_H = H-plane slant length
l_E = E-plane slant length
ϕ = flare angle in H plane
θ = flare angle in E plane
w_H = H-plane wave-guide width
w_E = E-plane wave-guide width

Note that all these dimensions, and the rest in the design section, are inside dimensions (ID) and are expressed in wavelengths.

To accomplish a union between the flared horn section and the wave guide at a common plane, the H-plane dimension in this procedure has been altered slightly from the optimum dimension. This reduces the directivity in the H plane and the over-all directivity. The reduction is slight, however, and can be taken into account by using a directivity 4 to 5 per cent higher than that desired.

The axial length L of this horn from the plane of the junction with the wave guide to the plane of the aperture is

$$L = \frac{b - w_E}{b}\sqrt{l_E^2 - \left(\frac{b}{2}\right)^2}$$

or

$$L = \frac{a - w_H}{a}\sqrt{l_H^2 - \left(\frac{a}{2}\right)^2}$$

The length along a common edge R between wave-guide junction and aperture is

$$R = \frac{b - w_E}{b}\sqrt{l_E^2 + \left(\frac{a - w_H}{2}\right)^2}$$

[1] See reference 15.
[2] See reference 331.

The slant length of the horn section along l_H between the aperture and the wave-guide junction is

$$\frac{b - w_E}{b} \sqrt{l_E^2 - \left(\frac{b}{2}\right)^2 + \left(\frac{a - w_H}{2}\right)^2}$$

and along l_E is

$$\frac{a - w_H}{a} \sqrt{l_H^2 - \left(\frac{a}{2}\right)^2 + \left(\frac{b - w_E}{2}\right)^2}$$

A procedure for designing *optimum* sectoral horns can be evolved as follows. Consider first an H-plane horn. The directivity of an optimum H-plane sectoral horn is

$$g_H = 7.86ab$$

where a = H-plane aperture dimensions
b = E-plane aperture dimensions = w_E (all dimensions are ID in wavelengths)
If the directivity is specified,

$$a = g_H/7.86w_E$$

The path length δ_H for an optimum H-plane sectoral horn is 0.397. Therefore,

$$\delta \doteq 0.397 \doteq a^2/8l_H$$

or

$$l_H \doteq a^2/3.18$$

The axial length between the plane of the junction of the wave guide and the plane of the aperture is

$$L_H = \frac{a - w_H}{a} \sqrt{l_H^2 - \left(\frac{a}{2}\right)^2}$$

Likewise, the directivity of an *optimum* E-plane sectoral horn is

$$g_E = 8.0ab$$

where a = H-plane aperture dimension = w_H
b = E-plane aperture dimension (all dimensions are ID in wavelengths)
If the directivity is specified,

$$b = g_E/8w_H$$

For an E-plane optimum horn $\delta_E \doteq 0.25$; therefore,

$$\delta_E \doteq 0.25 \doteq b^2/8l_E$$

or

$$l_E \doteq b^2/2.0$$

The axial length between the plane of the junction with the wave guide and the plane of the aperture is

$$L_E = \frac{b - w_E}{b} \sqrt{l_E^2 - \left(\frac{b}{2}\right)^2}$$

The shape (side-lobe level and beam width) of the E-plane and/or H-plane patterns of these antennas can be obtained by reference to Figs. 73 and 74.

When small horns are used as feeds for paraboloidal reflectors, an important characteristic is the 10-db (tenth-power) beam width of the radiation pattern. Empirical formulas that can be used as a good first approximation in the design of horn feeds are (the average flare angle is about 20 deg)[1]

$$H\text{-plane 10-db BW} = 31 + 79/a \quad \text{deg } (a < 3.0)$$
$$E\text{-plane 10-db BW} = 88/b \quad \text{deg } (b < 2.5)$$

where a = H-plane aperture dimension
b = E-plane aperture dimension (all dimensions are ID in wavelengths)

[1] See reference 15.

75. Parabolic-reflector Antennas.[1] Narrow-beam antennas require a radiating aperture which is very large compared with the wavelength. The parabolic-reflector antenna is a simple and convenient device to achieve this. It will operate over a relatively large band of frequencies; the band width is usually determined by the characteristics of the primary feed and the reflections from the surface back into the feed.

In principle, the parabolic-reflector antenna has the property of producing a plane wave front over its aperture from a point source located at its focus. This can be shown as follows. To produce a plane wave front from the reflecting surface shown in Fig. 78, the electrical distance from the point source to a plane perpendicular to the axis of the reflector along any two paths must be equal. For example, the distance from the point source at F to O and back must equal the distance FAB, that is

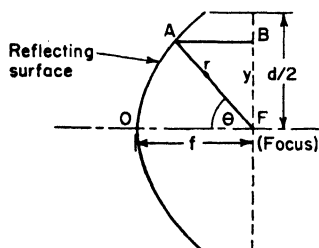



FIG. 78. Dimensions of the parabolic-reflector antenna.

$$2f = r(1 + \cos\theta)$$

or

$$r = 2f/(1 + \cos\theta)$$

This is the standard equation of a parabola in polar coordinates with its focus at the origin where f is defined as the focal length. In cartesian coordinates the standard equation with the origin at the vertex is $y^2 = 4fx$. Hence, the reflecting surface must be parabolic in shape.

A parabolic-cylinder reflector converts cylindrical waves from a line source at the focus into plane waves at the aperture. (Special types of parabolic-cylinder reflectors are used where accurately shaped patterns are required.) Antennas of this type are generally used where maximum directivity from an aperture is required and have been extensively employed in radars, both air-borne and ground-based, and recently in tropospheric-scatter communications systems and for radio astronomy. To achieve high directivity and maximum efficiency from the antenna, close control of the radiation from the primary feed illuminating the surface is required. That is, the primary feed must be small, produce spherical phase fronts, adequately illuminate the entire reflector surface without excess "spill-over," and produce such a field that after reflection the waves will be properly polarized. Some of these requirements are contradictory and compromises have to be made.

If the phase fronts of the radiation from the primary feed are not spherical, the wave fronts after reflection are not plane and the radiation pattern will be distorted and the directivity of the antenna reduced. The main effect of phase errors over the aperture, with common aperture distributions, is, in general, to reduce the gain and broaden the main lobe, particularly at the lower levels. The side-lobe level may be either raised or lowered depending upon the type of phase distortion and the intensity (amplitude) distribution over the aperture. Generally the nulls between the side lobes are filled in and their levels raised. Severe phase errors may result in splitting of the main lobe and a large increase in the side-lobe level.

To make the most effective use of the aperture of the parabolic-reflector antenna (*i.e.*, to obtain maximum directivity), the energy radiated from the primary source must be distributed uniformly over the surface. Practically, however, this is impossible without loss of energy due to "spill-over." Energy not intercepted by the reflector contributes to the side-lobe region and reduces the directivity of the point-source reflector-antenna system. Likewise, if the aperture illumination is tapered down in magnitude toward the edge to zero to reduce "spill-over," the directivity of the antenna again decreases. Therefore, somewhere between the uniform amplitude illumination and the zero-edge taper is an optimum amplitude distribution for maximum directivity. The ratio of the directivity of the parabolic-reflector antenna to the directivity of a uniformly illuminated aperture of the same area, which has the highest

[1] See references 9, 15, 341, and 360.

gain for uniform phase distribution, can be used to express the efficiency of the antenna illumination. This ratio is the directivity of the parabolic reflector relative to a uniformly illuminated aperture of the same area and depends upon the shape of the radiation pattern of the primary feed. It has been found experimentally that maximum directivity can be obtained for a circular paraboloid when the beam width to the −10-db points on the primary pattern corresponds to the angle subtended by the reflector at the focus. It is not critical, however. For a well-designed antenna, this maximum directivity is approximately 65 per cent of that of a uniformly illuminated circular aperture. To state it in another way, the effective aperture of a well-designed paraboloidal-reflector antenna is about 65 per cent of the projected area of the reflector. The incident energy at the edge of the reflection in this case is further reduced because of the increased space attenuation in the longer path. Figure 79 illustrates the tapering of the aperture illumination with an isotropic point, or line, source at the focus and can be used to determine the edge illumination.

Fig. 79. Aperture illumination of the parabolic-reflector antenna due to an isotropic radiator at the focus.

From the above considerations an order of magnitude of the directivity of a large practical paraboloidal reflector with the optimum amplitude illumination can be obtained as follows. The directivity of a large uniformly illuminated aperture is $4\pi \times \text{area}/\lambda^2$. For a large circular aperture of diameter d, this becomes $(\pi d/\lambda)^2$, and so the maximum directivity g of a large paraboloidal reflector is

$$g \doteq 6(d/\lambda)^2$$

The radiation pattern of a uniformly illuminated circular aperture has a half-power beam width of approximately $58\lambda/d$ deg with the first (and highest) side lobe 17.5 db below the main lobe. Beam widths of practical paraboloidal reflectors have been determined empirically (to within 10 per cent) and are shown below.[1]

$$\text{10-db beam width} = 130\lambda/d$$
$$\text{6-db beam width} = 100\lambda/d$$
$$\text{3-db beam width} = 70\lambda/d$$

The side-lobe levels that would be predicted from the diffraction pattern of the aperture of the antenna are often less than the direct radiation from the primary feed and the scattering from inaccurate reflectors and supporting structures used in practice.[2] The surface tolerance depends greatly upon the application. It has been found that for normal purposes a deviation of $\pm\lambda/8$ from the true contour is permissible. However, if the reflector is to be designed to produce low side lobes or a shaped beam,

[1] See reference 360.
[2] See references 334 and 347.

then a deviation of about $\pm\lambda/30$ is required. A tolerance of less than $\lambda/16$ at 8.5-mm wavelength (*i.e.*, 0.010 in.) from the true contour was held in the construction of the 50-ft "Big Dish" at the Naval Research Laboratory for use in radio astronomy.[1]

Feed lines to the primary radiator and support bars should run parallel to the H lines of the field; *i.e.*, there should be no component of E parallel to metal if possible. The best way to determine the effect of supports is to measure the directivity with and without them.

A simple paraboloidal-reflector antenna can be designed using the preceding information as follows. If the gain is specified, the diameter d of the reflector can be determined from the above considerations. The focal length can be determined from the 10-db beam width, $2\theta_{10}$, of the primary pattern and the equation of the parabola in polar form. That is, $f = [d(1 + \cos\theta_{10})]/[4 \sin\theta_{10}]$. If required, the diffraction pattern of the aperture can be calculated from the actual aperture distribution, which can be obtained from the primary pattern and Fig. 79. If low side lobes are important a greater amplitude taper can be used, or if high directivity is most important a more uniform illumination can be used.

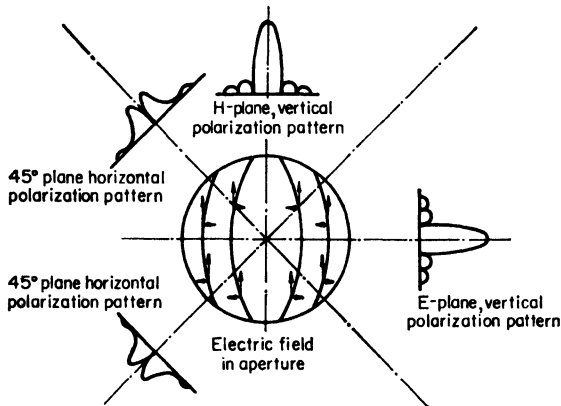



Fig. 80. Electric field in the parabolic-reflector aperture and resulting far-zone radiation patterns when the paraboloid is excited by a vertically oriented electric dipole.

76. Practical Primary Source Feeds for Parabolic-reflector Antennas.[2] The general requirements of a feed for a paraboloidal-reflector antenna are, as stated previously: (1) the feed must be small so that it appears as a point source, and to avoid excessive scattering from it, (2) it should produce spherical equiphase surfaces, (3) it must have the proper polarization characteristics, and (4) it should efficiently illuminate the entire reflecting surface. The shape of the pattern from it is limited, however, by the necessity to keep it small.

In addition, the input impedance of the feed must match that of the transmission line over as broad a band of frequencies as possible. The limits are determined by the application. For example, the impedance band width of a transmitting antenna is usually taken to be the band of frequencies over which the SWR is less than 1.5. The power-handling capacity of the feed is another important consideration. This is limited by the electrical breakdown between component parts and is of particular significance in aircraft antennas because the decrease in atmospheric pressure and the increase of free-ion content cause a decrease in breakdown potential with altitude. Therefore, feeds for aircraft antennas must be pressurized. The feed should be mechanically strong yet lightweight, and its design must specify reasonable tolerances so production methods can be used in its construction.

Both front and rear feeds are used. Front feeds are simpler electrically; rear feeds are simpler mechanically. Offset feeds are sometimes employed to remove the feed from the high-intensity fields.

The polarization characteristics of the primary radiator should be such that the waves will be polarized in the same direction after reflection from the paraboloid surface. Field components which are polarized perpendicular to the average are wasted

[1] See reference 350.
[2] See references 15, 341, and 360.

and contribute to side-lobe radiation. It has been shown theoretically that either electric-dipole- or magnetic-dipole-type feeds produce a field configuration in the aperture of a paraboloidal reflector that results in cross-polarized side lobes with a maximum in the planes at 45 deg to the principal planes,[1] as shown in Figs. 80 and 81. If a plane-wave field is used (*i.e.*, the proper combination of electric and magnetic dipoles), the aperture field configuration is such that the radiated field has only one field component, as shown in Fig. 82. In this case the E- and H-plane patterns of the feed are equal and patterns in the other planes are very similar, and therefore the

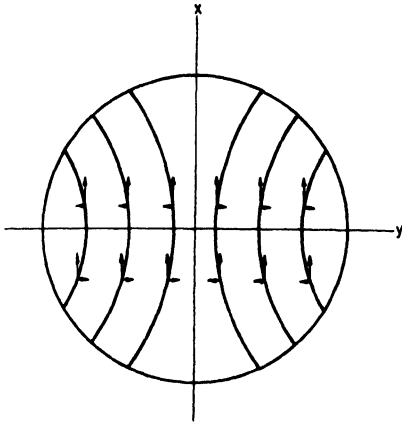



Fig. 81. Electric field in paraboloid-reflector aperture when paraboloid is excited by a short magnetic dipole lying along the y axis.

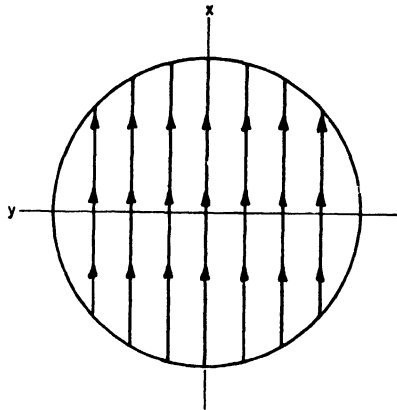



Fig. 82. Electric field in paraboloid-reflector aperture when paraboloid is excited by a small plane-wave source polarized along x axis.

main beam from the reflector has nearly a circular cross section. In practice, a horn-feed approximation to a plane-wave feed produced cross-polarized side lobes in the planes 45 deg to the principal planes at least 25 db below the principal-plane maximum.

77. Feed Systems. The most commonly used rear feeds are electric dipoles with parasitically excited reflectors, the dual-aperture (Cutler) feed, and the two-dipole wave-guide feed. Small horns, including open-ended wave guides, are generally used as front feeds.

78. Dipole Feed. Electric-dipole feeds are generally employed at frequencies where coaxial transmission lines are used. Electric dipoles used without reflectors are not suitable feeds because the pattern is uniform in the plane perpendicular to its length (H plane) but nonuniform ("figure-eight" pattern) in the planes that include its length; hence half or more of its radiated energy is not directed toward the surface. To increase the directivity of the primary pattern, dipoles are used with tuned-wire reflectors, plane-sheet (disk), half-cylinder, and hemispherical reflectors. For example, a dipole and a parasitically excited element tuned to a reactance of about 30 ohms and located $\lambda/8$ behind the driven element to act as a reflector are commonly used. A reflecting disk at least $\lambda/2$ in diameter is used with spacings up to 0.4λ. Since the spacing in this case is not too critical it is used in matching the dipole to the feed line. Even

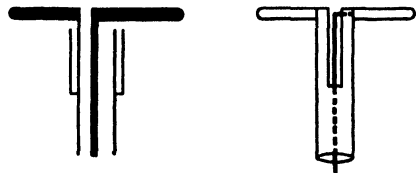



Fig. 83. Two methods of exciting balanced currents on a dipole feed.

[1] See references 341 and 353.

when used with a reflector, dipole feeds are limited by the range of patterns that can be obtained. Another disadvantage is the problem of feeding the dipole with balanced currents from an unbalanced (coaxial) line without producing currents on the outside of the outer conductor which cause spurious radiation. Various types of baluns (balance-to-unbalance transformers) can be used to overcome this effect. Two methods commonly used are the *quarter-wave-length skirt* and the *slot-balancing stub*, which are shown in Fig. 83.[1]

The power-handling requirements must also be considered in the design of dipole feeds. The parts of the dipole must be sufficiently spaced to prevent arc-over, and sharp corners must be avoided. Radiation from widely spaced slot-balancing systems seriously distorts the pattern of the feed, and shielding by a metal can is required. In general, design data for dipole feeds must be obtained experimentally because of the difficulty of predicting the properties theoretically.

Fig. 84. Wide-band dipole for the 50- to 60-cm band.

To obtain minimum cross-polarization effects and the best aperture distribution using a dipole-type feed, it has been found that the reflector should subtend an angle of about 140 deg at the focus.

79. Matching Problems. Two factors contribute to the mismatch at the input terminals of any feed: (1) the mismatch caused by the feed itself and (2) reflections from the reflecting surface of the antenna back into the feed. These are usually treated separately. The impedance of dipole-type feeds cannot be predicted with any degree of accuracy; so variation in impedance with frequency must be determined from measurements. Narrow-band impedance matching consists in varying the length of the driven dipole to tune out the reactance, and then using a $\lambda/4$ transformer to match the feed line. Broad-band matching at the higher frequencies has been accomplished by using thick dipole elements (tapered at the feed end to reduce capacitance) since the change in reactance with frequency is less for thick dipoles. Stub compensation can also be used.

At the lower frequencies, where dipole feeds are used almost exclusively, stub compensation is used to increase the band width.[1] This is accomplished as follows. First the dipole length is adjusted until its conductance at mid-band is equal to that of the feed line, and then the susceptance is removed with a shorted stub in parallel with the dipole, as shown in Fig. 84. Then, as the frequency changes, the susceptances of the dipole and stub vary in opposite directions. By adjusting the characteristic impedance of the stub and its length (by adding half wavelengths) it is possible to get good compensation over a frequency band.[2]

80. Wave-guide Feed. At centimeter wavelengths wave-guide feeds are practical. The two-dipole wave-guide-type feed shown in Fig. 85 has been extensively used because it is much simpler to feed a dipole from a wave guide than from a coaxial line.[3] The dipoles are mounted on a web that fits into the mouth of the wave guide so the dipoles are parallel to the electric field of the dominant $TE_{1,0}$ mode. Equal excitation is achieved if the web is symmetrically located. The combination is tuned to reflect

[1] See reference 360.
[2] See reference 372.
[3] See reference 15, p. 255.

power back into the reflecting surface. The closest dipole is resonant; the second dipole acts as a parasitic element. The second dipole can be replaced by a disk. The wave guide is tapered to serve as an impedance-matching device. It also improves the radiation pattern because it helps decouple the outer wall from the dipole. The impedance of the antenna is also determined by the depth of the web and the position of the dipoles with respect to the mouth.

Another wave-guide feed is the *dual-aperture rear feed* (*Cutler feed*)[1] which essentially splits the energy in the wave guide into two branches that turn back on themselves

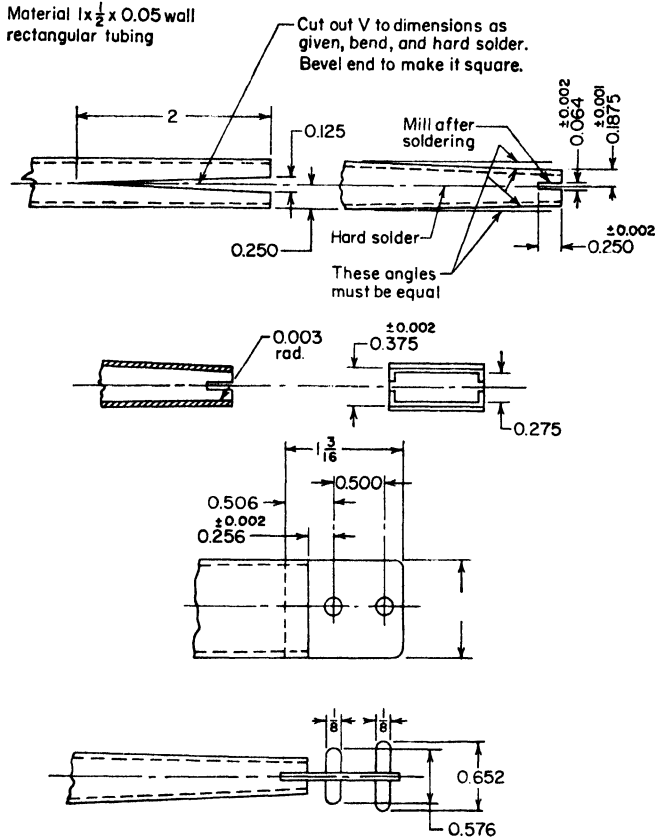



Fig. 85. Wave-guide couple-dipole feed; $\lambda = 3.2$ cm.

and radiate toward the reflector through two slots. A sketch is shown in Fig. 86. It is important that the two apertures be less than $\lambda/2$ apart and that the apertures be not too close to the wave-guide wall. Therefore, the wave guide must be tapered to something less than $\lambda/4$ in the narrow dimension. The two branch paths are contained in a cylindrical cavity. The dimensions of the cavity and the indentation at its center can be used in matching the feed to the wave guide. This is usually determined by experiment. Band widths of 6 per cent have been obtained at X band for an SWR less than 1.3. The phase front on either side appears to be coming from the nearer aperture and its image to be reflected in the wave-guide wall. This results in phases in the E and H planes at the edge of the reflector which differ by about 30 deg.

[1] See reference 341.

The shape of the pattern in the *E* plane can be controlled by the separation of the slot from the wave-guide wall and somewhat by slot width. The *H*-plane pattern can be controlled by the length of the slot.

Recently, a wave-guide feed having equal *E*- and *H*-plane radiation patterns has been developed.[1] The feed consists of an electric-dipole source and a parasitically excited open-ended coaxial line carrying a $TE_{1,1}$ mode which is similar to a magnetic-dipole source. These two sources have complementary pattern characteristics and are combined to compensate each other so equal beam widths can be obtained. It is constructed from a rectangular wave guide tapered in the narrow dimension, as shown in Fig. 87. A hole is drilled into the broad face of the wave guide and a bead-supported dipole is centered in the hole. The $TE_{1,1}$ source is obtained by placing a metallic cylinder around the wave guide. The end of the cylinder farthest from the dipole is short-circuited. The dipole excites a $TE_{1,1}$ mode in this irregular coaxial line. The location and length of the cylinder are adjusted for equal beam widths in the *E* and *H* planes. The relative beam widths can be controlled by the length of the cylinder and the location of the short circuit relative to the dipole. Narrow beam widths are obtained for closer short-circuit spacings. Design data are shown in Fig. 88.

FIG. 86. Dual-aperture rear-feed horn.

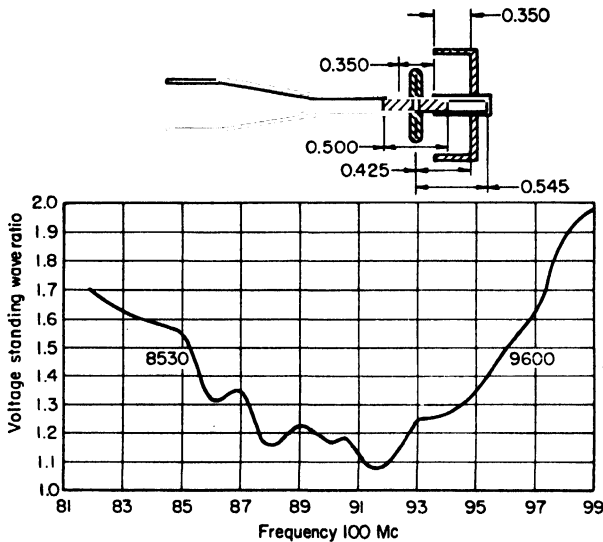



FIG. 87. Sketch of an antenna feed, and its performance at *X* band.

The impedance of the feed depends upon the position of the wave-guide short circuit beyond the dipole, the length and diameter of the dipole as well as its supporting Teflon bead, and the hole diameter through which the dipole protrudes. These parameters also determine breakdown power so the problem of adjusting parameters reduces itself to obtaining the largest power-handling capacity with the greatest band width. A table of these characteristics follows.

[1] See reference 335.

Hole diam, in.	Bead diam, in.	% band width (SWR less than 1.5)	Breakdown power, kw
0.450	0.500	2	120
0.350	0.500	11	75
0.250	0.500	22	50

81. Front-feed Systems. The most commonly used front feeds are open-ended wave guides and horn antennas. In general, the radiation characteristics of open-ended wave guides depend upon the size and shape of the aperture and the mode or modes propagating in the wave guide. It has been found that for a circular paraboloid a circular wave guide propagating a $TE_{1,1}$ mode has ideal phase and polarization characteristics with suitable directivity (a 10-db beam width of approximately 130 deg in both principal planes).[1] Greater directivity can be obtained by flaring the aperture

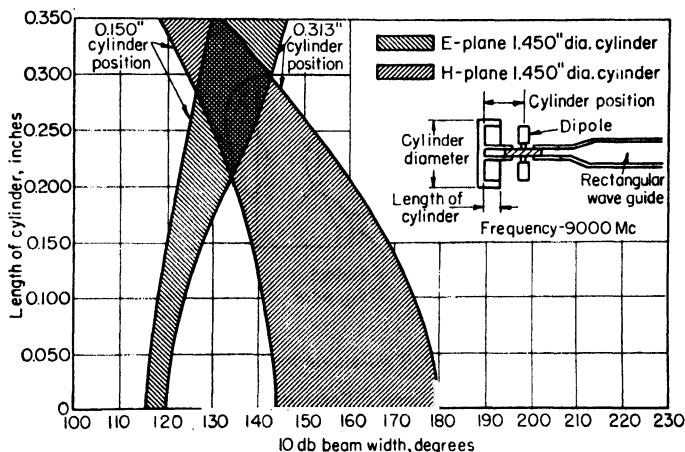



FIG. 88. Effect of cylinder length and position on the E- and H-plane beam widths.

into a small conical horn. An open-ended rectangular wave guide propagating a $TE_{1,0}$ mode does not give a circularly symmetric radiation pattern but is suitable for feeding a paraboloidal section which is cut to subtend a wide angle in the E plane and a narrow angle in the H plane. This is because the pattern in the E plane is quite broad with considerable side and back radiation; the H-plane pattern is much more directive with a half-power beam width of about 50 to 60 deg. The pattern shape can be controlled more or less independently by the corresponding aperture dimension. This feed has good phase characteristics and tolerable polarization characteristics. The shape of the E-plane pattern can be improved by the addition of flanges along its wide dimension. The flange can be about $\lambda/2$ wide. Alternatively, the H-plane pattern can be altered by changing the shape of the mouth of the guide.[2]

82. Wave-guide-horn Feed. Where more directivity is required, some form of wave-guide horn may be used.[3] In this case there are two important factors, the location of the center of phase in the E and H planes and the magnitude of the back radiation. For example, a sectoral horn sometimes used as a feed for elliptical paraboloid sections has poor phase characteristics because the phase center in the plane of the flare is near the apex of the flare angle while in the other plane the phase center is at

[1] See reference 341.
[2] See reference 329.
[3] See Sec. 87.

the horn aperture and these may be several wavelengths apart. Since horn feeds are not resonant, the input impedance is close to that of a uniform wave guide and its magnitude does not change rapidly with frequency. The SWR associated with open-ended wave guides of standard dimensions varies from about 1.8 to 3.0 depending upon the aperture size in wavelengths.

A *two-mode horn* or *box-horn feed*[1] has been designed for elliptical paraboloid sections. Two modes $TE_{1,0}$ and $TE_{3,0}$ are excited in an open-ended wave guide by symmetrically terminating a normal-size wave guide in a guide wide enough to support both modes, as shown in Fig. 89. The relative strength of the two modes is controlled by the ratio of the wave-guide widths and the relative phase is controlled by the length of the wide

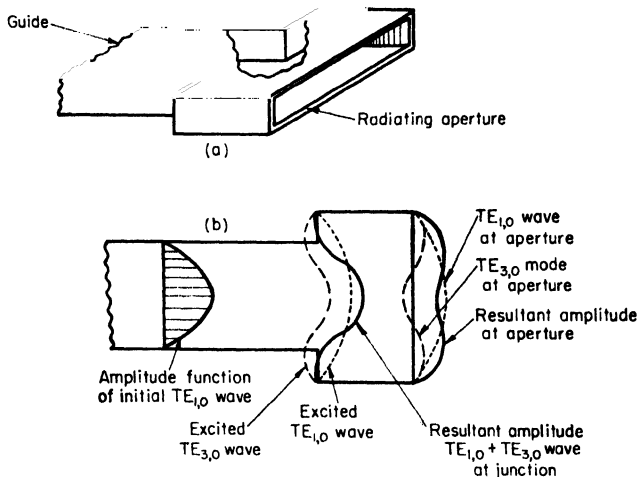



FIG. 89. Two-anode feed horn.

wave guide. When the relative phase differs by 180 deg, a resultant in-phase field exists across the aperture with almost constant magnitude. In practice these conditions hold for approximately a 10 per cent band of frequencies.

83. Hoghorn Feed. A wave-guide feed that combines the properties of a horn and parabolic reflector is shown in Fig. 90. It is commonly called the hoghorn feed.[2] This type of feed, and its variations, is a compact electromagnetic horn that has a rectangular aperture which may be from 4λ to 15λ long and from 2λ to 3λ wide and produces a fan-shaped beam of from 4 to 15 deg by approximately 25 deg. It is used as a primary feed for a parabolic-cylinder reflector with closed ends (commonly called the *cheese antenna*). Construction details are shown in Fig. 91. The parabolic surface can be obtained from

$$f = 0.70(a - 0.933w)$$

where f = focal length
$a = BC$
w = wave-guide width ED

The E-plane radiation patterns are essentially the same as a sectoral horn of the same size. The pattern is not quite symmetrical in the H plane with a 17-db side lobe on one side only. The direction of maximum radiation is approximately parallel to the parabola axis. The side-lobe level is critically dependent upon the shape of the horn near the focus.

84. Mismatch Due to Feed. In practice, the presence of the feed in the field has a detrimental effect on the characteristics of the reflector antenna, especially if the

[1] See reference 341.
[2] See reference 366.

antenna is to operate over a broad band of frequencies.[1] As stated above, the mismatch at the input terminals of any feed is due to the mismatch caused by the feed itself and the scattering from the surface of the reflector back into the feed. For example, the SWR on the feed line of a matched horn increased to 1.4 when placed at the focus of a reflector. The magnitude of the impedance of a well-matched feed is fairly constant but the phase varies rapidly with frequency because of the long path lengths and can cause instability of the transmitting oscillator. This can be overcome by using a circularly polarized feed (the wave reflected from the parabola is polarized in the opposite sense), by using an offset feed, or by using an *apex-matching plate*.[2] That is, a plate of dielectric is placed at the apex of the paraboloid to produce a signal equal in magnitude but 180 deg out of phase with that scattered from the remainder of the reflector. Since the sources of reflection are nearly the same distance from the focus, the impedance correction is effective over a wide band of frequencies. However, the directivity is decreased and the side-lobe level is increased because of scattering from the plate.

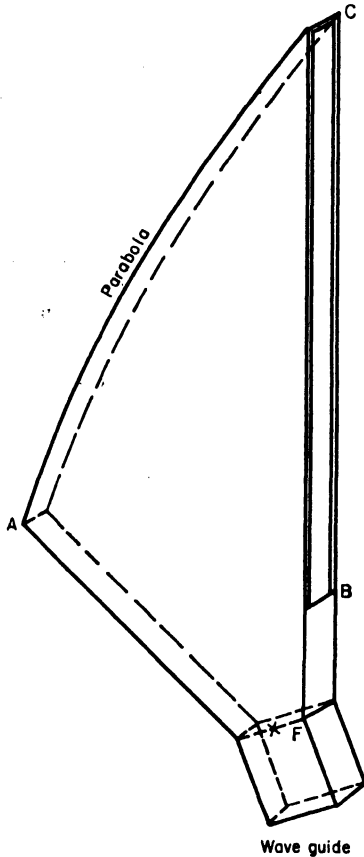



FIG. 90. Simple hoghorn.

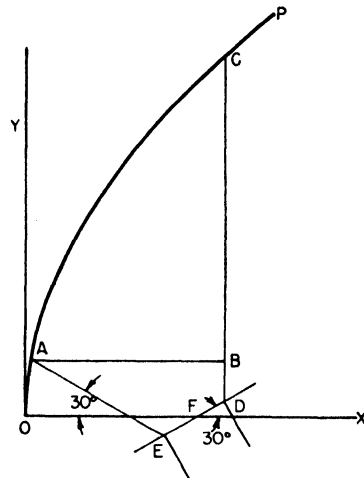



FIG. 91. Geometry of hoghorn.

Direct radiation from the feed to the sides and rear can change the directivity of the reflector antenna with frequency, especially with small paraboloids. In such cases the focal length of the parabola should be coordinated with the wavelength to ensure proper phase of rearward radiation. Even in the absence of direct-feed radiation the presence of the feeds has a detrimental effect on the radiation pattern because of aperture blocking and scattering from the surfaces of the feed. The effect is to reduce directivity and increase the level of the side lobes.

Beam scanning with only a small amount of defocusing can be produced by moving the feed laterally away from the focus. For example, the beam can be shifted about

[1] See reference 347.
[2] See reference 307.

twice the half-power beam width with only a 0.5-db decrease in directivity. The beam shift can be doubled if the feed is fixed and the reflector tilted.

85. Lens Antennas. At centimeter wavelengths, optical devices such as lenses may be applied. Since the fundamental function of a lens is to modify the phase fronts from a radiating source, the lens surfaces are designed using the laws of geometrical optics, as in the case of the parabolic reflector. However, the radiation pattern of the antenna must be considered from the standpoint of diffraction theory.

Because of the difference in wavelengths, microwave-lens techniques are free of certain restrictions which apply in the design of optical lenses. Surface tolerances are large; dielectric lenses can be made of relatively soft low-loss plastics; and the lens surfaces can be molded or turned on a lathe. Consequently, the surface need not be spherical but can be cut to contours appropriate to the function of the lens. Metal-plate lenses, likewise, can be produced by ordinary machine-shop methods.[1]

FIG. 92. Dimensions of dielectric-lens antenna.

Lens antennas can be divided into two types: (1) those in which the electrical (optical) path length is increased by the lens medium (called *delay lenses*), such as natural-dielectric and H-plane metal-plate lenses, and (2) those in which the electrical path length is decreased by the medium, such as the E-plane metal-plate lens.

Dielectric lenses may in turn be divided into three groups according to the materials from which they are constructed: (1) lenses of homogeneous isotropic nonmetallic dielectric material, *e.g.*, polystyrene; (2) lenses of dielectric materials of variable or nonuniform dielectric constant (such as the *Luneberg lens*); and (3) lenses of artificial dielectrics.

86. Dielectric-lens Antennas. Lens design is based on two principles of geometrical optics: (1) the principle of equality of electrical (optical) paths along rays between pairs of wave fronts; and (2) Snell's law of refraction. The design procedure is called ray tracing.

To straighten a spherical wave front, the contour of the lens shown in Fig. 92 is

$$R = \frac{(n-1)f}{n\cos\theta - 1}$$

where f = focal length of lens

n = index of refraction of dielectric medium

$$= \sqrt{\frac{\mu\epsilon}{\mu_0\epsilon_0}} = \sqrt{\mu_r\epsilon_r}$$

If $\mu = \mu_0$, $n = \sqrt{\epsilon_r}$, the square root of the relative dielectric constant. In rectangular coordinates the equation of the surface contour is

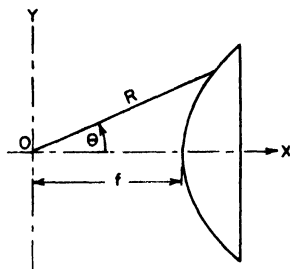
$$\left[\frac{(n+1)x}{f} - n\right]^2 - \left(\frac{n+1}{n-1}\right)\frac{y^2}{f^2} = 1$$

which is the equation of a hyperbola. If the origin is moved to the vertex, the equation is

$$\left[\frac{(n+1)x}{f} + 1\right]^2 - \left(\frac{n+1}{n-1}\right)\frac{y^2}{f^2} = 1$$

The asymptotic angle θ_a is given by $\cos\theta_a = 1/n$. Since the right-hand surface lens surface is coincident with an equiphase surface it has no effect on the direction of the rays.

For a point-source feed, the lens surface is a spherical hyperbola, and for a line source feed, it is a cylindrical hyperbola. If the angle θ is small, the shape is approximately spherical.

[1] See reference 15.

Neglecting the attenuation in the material and power loss caused by reflections at the surfaces, the power density $P(\rho)$ watts/m², at any point in the aperture of a spherical hyperbolic lens at a distance ρ ($= R \sin \theta$) from the axis, is related to the geometry of the lens and the power pattern $P(\theta)$ of the point-source primary feed (the primary feed is considered a point source because it should be very much smaller than the lens) by

$$P(\rho) = \frac{(n \cos \theta - 1)^3}{(n - 1)^2(n - \cos \theta)} P(\theta)$$

where n = index of refraction.

The corresponding amplitude $A(\rho)$ at any point in the aperture is

$$A(\rho) = \sqrt{P(\rho)} = \frac{1}{n - 1} \sqrt{\frac{(n \cos \theta - 1)^3}{n - \cos \theta}} A(\theta)$$

where $A(\theta)$ = primary field-strength pattern.

For a cylindrical hyperbolic lens, the power density $P(y)$ watts/m² at any point y ($= R \sin \theta$) from the lens axis, in a cross-sectional plane, is

$$P(y) = \frac{(n \cos \theta - 1)^2}{(n - 1)(n - \cos \theta)} P'(\theta)$$

where $P'(\theta)$ = primary power pattern of the line-source feed as a function of θ in the cross-sectional plane.

The corresponding amplitude $A(y)$ at any point y in the aperture is

$$A(y) = \sqrt{P(y)} = \frac{n \cos \theta - 1}{\sqrt{(n - 1)(n - \cos \theta)}} A'(\theta)$$

Hence, the aperture distribution for a given lens can be determined and the radiation pattern calculated when the primary pattern of the feed is known.

The amplitude functions given above drop off rapidly with θ. For example, for a "spherical" lens of polystyrene ($n = 1.6$), the amplitude has dropped off 50 per cent (6 db) at angles 30 deg from the axis. Larger angles than this result in a high degree of taper to the illumination and a consequent reduction in directivity relative to a uniformly illuminated aperture. For a nearly uniform illumination, therefore, the angle θ must be kept fairly small.

The most serious disadvantage of dielectric lenses of conventional design is their large mass. The dimension of the simple lens described above can be obtained as follows. If (R_m, θ_m) are the polar coordinates of the edge of the lens and t is the lens thickness on the axis, then from the principle of equality of electrical paths

$$t = \frac{R_m(1 - \cos \theta_m)}{n - 1}$$

where n = index of refraction.

A lightweight short-focal-length lens can be obtained by the process of *zoning*, which results in the cross section of Fig. 93. The thickness of a zoned step z is such that the electrical length of z in the dielectric is an integral number of wavelengths (usually taken as one) longer than the electrical length in air. That is, $z/\lambda_d - z/\lambda_0 = 1$, or

$$z = \frac{\lambda_0}{n - 1}$$

where λ_d = wavelength measured in dielectric

λ_0 = wavelength measured in air

Hence, the thickness of a zoned lens will be slightly larger than $\lambda_0/(n - 1)$.

The zoned lens is frequency-sensitive. An approximate expression for the band

width for which the phase variations in the wave front from a plane do not exceed $\lambda/8$ is

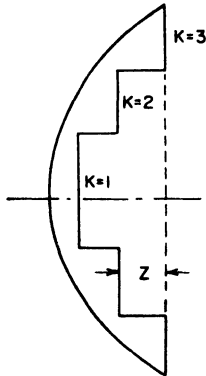



Fig. 93. Zoned-lens antenna.

$$\text{Band width} \doteq \frac{25}{K - 1} \quad \text{per cent}$$

where K = number of zones. The zone on the axis is counted as the first zone.

Reflections from dielectric surfaces can cause feed mismatch and power instability and are most likely to occur when a lens surface coincides with an equiphase surface, *e.g.*, the right-hand surface of the simple dielectric lens. The surface-reflection coefficient R at normal incidence is

$$R = -\frac{n - 1}{n + 1}$$

The magnitude of the reflection coefficient increases with n. The lens may be tilted slightly so that the reflected wave comes to focus to one side of the feed without seriously affecting the performance of the lens.

The magnitude of the reflection coefficient also depends upon the angle of incidence. The power loss caused by these reflections reduces the antenna gain and increases the side-lobe level. Figure 94 gives the fraction of the incident power $(R)^2$ reflected from the surface of an infinite slab of dielectric ($n = 1.6$) as a function of the angle of incidence for polarization (1) perpendicular to the plane of incidence and (2) parallel to the plane of incidence. A $\lambda/4$ matching surface of refractive index $\sqrt{n}$ can be placed over the reflecting surfaces to reduce reflections.

Fig. 94. Fraction of incident power reflected from the surface of an infinite dielectric slab vs. angle of incidence. (*a*) Electric vector perpendicular to the plane of incidence; (*b*) electric vector in the plane of incidence.

87. Artificial-dielectric Lens Antennas.[1] Artificial-dielectric materials can be made up from an array of discrete metal particles or obstacles simulating, as it were, molecular particles supported in a matrix of strong yet lightweight dielectric. The increase in path length (or delay in phase velocity) produced by the metallic particles is due to the increased average polarization of the dielectric medium. The size of the metal particles should be small compared with the wavelength to avoid resonance. (The maximum length parallel to the E field should be less than $\lambda/4$.) The spacing should be less than a wavelength to avoid diffraction effects. The advantages of artificial-dielectric materials over conventional dielectrics include light weight, mechanical strength, and a wide range of refractive indices.

The effective relative dielectric constant and permeability of a number of particles are given on page 20–109.[2] The formulas listed are valid for the particular orienta-

[1] See references 336, 337, 338, 346, 357, and 368.
[2] See reference 338.

Table of Relative Dielectric Constant and Permeability for Various Obstacle Shapes

Obstacle shape	Dielectric constant	Permeability
Thin circular disks	$\epsilon_r = 1 + \frac{2d^3N}{3}$	$\mu_r = 1 - \frac{d^3N}{3}$
Thin ellipses	$\epsilon_r = 1 + \frac{4\pi ab^2k^2N}{3(1-k^2)[F(k)-E(k)]}$, $\epsilon_r = 1 + \frac{4\pi ab^2k^2N}{3[E(k)-(1-k^2)F(k)]}$	$\mu_r = 1 - \frac{4\pi ab^2N}{3E(k)}$
Thin squares	$\epsilon_r = 1 + 1.032\,d^3N$	$\mu_r - 1 - 0.455\,d^3N$
Thin rectangles	$\epsilon_r = 1 + \frac{\pi w^2 lN}{4}$ ($l >> w$)	$\mu_r = 1 - \frac{\pi w^2 lN}{4}$
Thin infinite strips	$\epsilon_r = 1 + \frac{\pi w^2 n}{4}$	$\mu_r = 1 - \frac{\pi w^2 n}{4}$
Circular cylinders	$\epsilon_r = 1 + 2\pi a^2 n$	$\mu_r = 1 - 2\pi a^2 n$ (H perpendicular to axis); $\mu_r = 1 - \pi a^2 n$ (H parallel to axis)
Spheres	$\epsilon_r = 1 + 4\pi a^3N$	$\mu_r = 1 - 2\pi a^3N$
Moderately thick obstacles		$\mu_r = 1 - \tau N$

N is the number of obstacles per unit volume, $k = \sqrt{1-(b/a)^2}$, $F(k)$ and $E(k)$ are complete elliptical integrals, n is the number of obstacles per unit area perpendicular to strips or cylinders, τ is the volume occupied by an obstacle.

tion of the pertinent field vector indicated in the table. The relative permeability of an array of very thin obstacles parallel to the magnetic field is unity. In this table, N is the number of obstacles per unit volume, $k = [1 - (b/a)^2]^{1/2}$, $F(k)$ and $E(k)$ are complete elliptical integrals, n is the number of obstacles per unit area perpendicular to the strips or cylinders, and τ is the volume occupied by an obstacle. The index of refraction is

$$\sqrt{\epsilon_r \mu_r}$$

where ϵ_r = relative dielectric constant
μ_r = relative permeability

Artificial-dielectric materials can also be constructed in the form of an assembly of two or more dielectric materials such as a three-dimensional array of quartz dielectric spheres supported by a foamed dielectric matrix.[1] Or, alternatively, spherical voids, or cylindrical holes, can be arrayed in a matrix of high-dielectric-constant material.[2] Alternate layers of high- and low-dielectric-constant materials have been used to make up artificial dielectric.[3] The relative dielectric constant may be adjusted to any value between those of the component materials by selecting the proper ratio of slab thickness.

Artificial dielectrics can be used to good advantage as matching layers for lenses of homogeneous dielectrics.[4]

88. *E*-plane Metal-plate Lens Antenna.[5] This type of lens is made up of a stack of parallel-plate transmission lines propagating a $TE_{1,0}$ wave. That is, the metallic plates of the transmission line are parallel to the direction of the electric field. The phase velocity for a $TE_{1,0}$ wave in such a line is

$$v = \frac{c}{\sqrt{1 - \left(\frac{\lambda_0}{\lambda_c}\right)^2}} \quad \text{m/sec}$$

where c = free-space velocity (3×10^8 m/sec, approximately)
λ_0 = free-space wavelength
λ_c = cutoff wavelength = $2a$ (a = plate separation)

The index of refraction for this lens is

$$n = \frac{c}{v} = \sqrt{1 - \left(\frac{\lambda_0}{\lambda_c}\right)^2}$$

Note that n is less than 1; hence the electrical paths are reduced by the insertion of a metal-plate lens. Values of n lie between 0 and 0.86. The smallest practical value of n should be used in order to keep the thickness at a minimum. The minimum practical value of n has been found to be about 0.5.

Y
R
y
O E θ x X
f

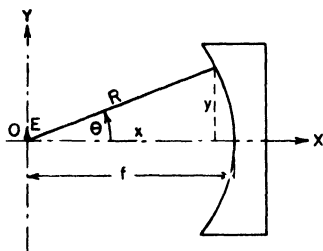



FIG. 95. Dimensions of an *E*-plane metal-plate lens antenna.

An *E*-plane metal-plate lens can be designed to have variable thickness and uniform spacing or uniform thickness and variable spacing. The contour of a lens with uniform plate spacing which will focus an incident plane wave is given by (see Fig. 95)

$$R = \frac{(1 - n)f}{1 - n \cos \theta}$$

where n = index of refraction.

[1] See reference 339.
[2] See references 346 and 368.
[3] See reference 338.
[4] See reference 362.
[5] See references 9, 13, 15, and 356.

In rectangular coordinates the equation is

$$\left[\frac{(1+n)x}{f} - n\right]^2 + \left(\frac{1+n}{1-n}\right)\frac{y^2}{f^2} = 1$$

This is the equation of an ellipse. For a point-source primary feed, the lens has a concave ellipsoidal surface and for a line source the lens has a concave cylindrical surface.

In the E plane, focusing is accomplished because of diffraction (Snell's law); however, in the H plane, the waves are constrained to travel parallel to the plates.[1]

A serious disadvantage of this type of lens is its limited bandwidth. If δ is the maximum tolerable path difference in wavelengths between the central and the outermost ray, and t is the thickness, in wavelengths, of the lens plate at the edge, then

$$\text{Band width} = \frac{2n}{1-n^2}\frac{\delta}{t}$$

The thickness of a zone step is (see Fig. 96)

$$z = \frac{\lambda_0}{1-n}$$

where λ_0 = free-space wavelength
n = index of refraction

The contour of the zoned lens is

$$R_K = \frac{(1-n)f_K}{1-n\cos\theta}$$

where $f_K = f + (K-1)z$
f = focal length of unzoned lens
K = number of zones = 1, 2, 3, . . . ; the central zone is counted as the first zone

Fig. 96. Cross section of a zoned E-plane metal-plate lens antenna.

Zoning an E-plane metal-plate lens increases its band width and decreases the size (thickness) of the lens. The band width is given approximately by

$$\text{Band width} \doteq \frac{50n}{1+Kn} \quad \text{per cent}$$

where n = index of refraction at the design frequency.

Practical lenses of this type have band widths of 4 to 5 per cent.

The amplitude $A(\rho)$ at any point in the aperture of a "spherical" lens a distance ρ ($= R\sin\theta$) from the lens axis is related to the geometry of the lens and the amplitude pattern $A(\theta)$ of the primary feed by

$$A(\rho) = \frac{1}{1-n}\sqrt{\frac{(1-n\cos\theta)^3}{\cos\theta - n}}\,A(\theta)$$

For a "cylindrical" lens, the amplitude $A(y)$ at any point y ($= R\sin\theta$) in a cross-sectional plane is given by

$$A(y) = \frac{1-n\cos\theta}{\sqrt{(1-n)(\cos\theta-n)}}\,A'(\theta)$$

where $A'(\theta)$ = field-strength pattern of the line-source feed in the cross-sectional plane as a function of θ.

The directivity is about the same as for an optimum horn of the same aperture.

[1] Constrained-type lens antennas are applicable to wide-angle scanning; see references 342, 359, 373, and 376.

The general problem of reflection from the surfaces of a metal-plate lens is quite complex. Formulas and tables for reflections of plane waves from the edge of an infinite array of parallel equally spaced plates when their edges lie in a plane can be found in the literature.[1] From these rough approximations the magnitude of the reflections from a practical antenna can be determined.

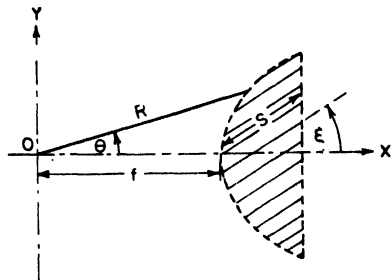



FIG. 97. *H*-plane metal-plate lens antenna.

89. *H*-plane Metal-plate Lens Antenna.[2] The velocity of a wave entering a stack of parallel-plate transmission lines oriented parallel to the direction of the magnetic field is essentially unchanged because the lines will support a *TEM* wave. However, the wave is *constrained* to travel between the plates, so that, once inside, the path length can be increased if the plates are deformed or slanted as shown in Fig. 97. In this type of lens

$$R = \frac{(n-1)f}{n\cos\theta - 1}$$

where n = effective index of refraction = $1/\cos\xi$ $(n > 1)$

ξ = slant angle of the parallel-plate transmission lines

In this lens the index of refraction, $1/\cos\xi$, is independent of frequency and depends only upon the slant angle ξ.

The most critical dimension is the slant length S. Assuming a maximum allowable variation of $\lambda/8$ in the electrical path length, the tolerance on S is

$$\Delta S = \pm 0.06\lambda$$

The chief disadvantage of this type of lens is that the aperture illumination is unsymmetrical in the E plane.

90. Tolerances on Lens Antennas.[3] The tolerances given in the table on page 20-113 are based on a maximum allowable deviation in path length (from all causes) of $\pm\lambda/16$ from a mean value (total variation of $\lambda/8$). A relatively large amount of warping and twisting can be tolerated in a lens antenna compared with a parabolic-reflector antenna, which is an advantage. The lens axis can be tilted with respect to the axis of the feed without serious pattern distortion.

91. Luneberg Lens Antenna.[4] Luneberg showed that, if a dielectric sphere has a variable index of refraction that varies with distance r from the center, as

$$n = \sqrt{2 - (r/a)^2}$$

FIG. 98. The Luneberg lens antenna.

where a = radius of the sphere, then all the energy entering the sphere at a point on the surface is focused into a beam directed diametrically opposite the point. This is shown in Fig. 98. The rays that leave the point source S on the lens surface are focused into parallel rays. A ray leaving the source at an angle θ is radiated from the lens at P so that the radius OP forms the angle θ. Because of symmetry, scanning in three dimensions can be achieved without pattern deterioration by moving the source on the lens surface.

[1] See references 15 and 333.
[2] See reference 358.
[3] See references 9 and 15.
[4] See references 361, 364, and 365.

The variation in the index of refraction is achieved by controlling the density of the dielectric material or by loading a low-dielectric constant material such as a foam with high-dielectric-constant materials.

The amplitude of the field $A(\rho)$ at any point in the projected aperture is related to the feed pattern $A(\theta)$ by

$$A(\rho) = \sqrt{\sec\theta}\, A(\theta) \qquad \text{for } -\frac{\pi}{2} < \theta < \frac{\pi}{2}$$

where θ = angle at which a ray leaves the source.

The restriction on θ is necessary since sec θ approaches infinity as θ approaches $\pm\pi/2$.

Tolerances on Lens and Reflector Antennas*

Type of antenna	Type of tolerance	Amount of tolerance
Parabolic reflector	Surface contour	$\pm 0.03\lambda_0$
Dielectric lens ($n > 1$)	Thickness	$\pm \frac{0.03\lambda_0}{n-1}$
	Index of refraction	$\pm \frac{3\lambda_0}{nt}$ %
Zoned dielectric lens ($n > 1$)	Thickness	± 3 %
	Index of refraction	$\pm \frac{3(n-1)}{n}$ %
E-plane metal-plate lens, unzoned ($n < 1$)	Thickness	$\pm \frac{0.03\lambda_0}{1-n}$
	Plate spacing	$\pm \frac{3n\lambda_0}{(1-n^2)t}$ %
E-plane metal-plate lens, zoned ($n < 1$)	Thickness	± 3 %
	Plate spacing	$\pm \frac{3n}{1+n}$ %

n = index of refraction
t = lens thickness
* See reference 9.

The half-power beam width for a Luneberg lens antenna fed from an open-ended wave guide or small horn is given approximately by

$$\text{Half-power BW} \doteq \frac{30\lambda_0}{a}$$

where a/λ_0 = lens radius in free-space wavelengths.

A two-dimensional Luneberg lens can be constructed from two circular disks forming a parallel-plate line. If the direction of the electric field is normal to the plates, the desired variation in the index of refraction can be achieved by using a slab of variable-index dielectric [$n = \sqrt{2 - (r/a)^2}$] placed between the plates.

For example, a simple method for obtaining a variable dielectric constant as a function of radial distance consists in drilling holes of gradually increasing diameter, as one progresses outward, into a slab of polystyrene. If the direction of the electric field is parallel to the plates, a *TE* wave will be propagated. The desirable index can be obtained by varying the plate separation since in this case, as in the E-plane metal-plate lens,

$$n = \sqrt{1 - (\lambda_0/\lambda_c)^2}$$

where λ_0 = free-space wavelength
λ_c = free-space cutoff wavelength = 2 × plate separation

The half-power beam width in the H plane of a two-dimensional Luneberg lens is given approximately by

$$H\text{-plane half-power BW} \doteq \frac{143}{\sqrt{a/\lambda_0}}$$

where a/λ_0 = lens radius, wavelengths.

92. Microwave Slot Antennas—Slotted-wave-guide Arrays.[1] If a narrow slot is cut into the wall of a wave guide so that its long dimension is in a direction transverse to the current lines, the internal wave-guide field is coupled to space. The degree of coupling depends upon the current density and the projected length of the slot normal to the current lines. This provides a method of excitation.

If the slot is cut so that its long dimension is parallel to the flow of current or is located where the current is zero, there is very little coupling of the internal field to space. Nonradiating slots provide means of entry into the wave guide for measuring the internal fields.

Narrow slots ($W \ll \lambda$) approximately $\lambda/2$ long are used as practical radiating elements on slotted-wave-guide arrays. In this case, the slot field is nearly sinusoidal along its length independent of the method of excitation.

FIG. 99. Parameters and equivalent circuits of slots in rectangular wave guide (reference point for circuit elements is the center of the slot): (*a*) wave-guide dimensions; (*b*) longitudinal slot in broad face, shunt element; (*c*) transverse slot in broad face, series element; (*d*) centered inclined slot in broad face, series elements; (*e*) centered inclined slot in narrow face, shunt element.

The simple equivalent circuits for the most commonly used resonant slots in a rectangular wave guide propagating the dominant $TE_{1,0}$ mode and the corresponding value of loading are given below. If g represents the shunt conductance of a slot normalized to the characteristic admittance of the $TE_{1,0}$-mode line, and r represents the series resistance of a slot normalized to the characteristic impedance, then for a longitudinal slot in the broad face (Fig. 99*b*),

$$g = 2.09 \frac{\lambda_g}{\lambda_0} \frac{a}{b} \cos^2\left(\frac{\pi\lambda_0}{2\lambda_g}\right) \sin^2\left(\frac{\pi x_1}{a}\right)$$

where λ_g = wave-guide wavelength.

The frequency characteristics of longitudinal slots is such that (for slots up to $\lambda/4$ wide) the wider the slot the flatter the response. For a transverse slot in the broad face (Fig. 99*c*), the series resistance is

$$r = 0.523 \left(\frac{\lambda_g}{\lambda_0}\right)^3 \frac{\lambda_0^2}{ab} \cos^2\left(\frac{\pi\lambda_0}{4a}\right) \cos^2\left(\frac{\pi x_1}{a}\right)$$

For a centered inclined slot (Fig. 99*d*), the series resistance is

$$r = 0.131 \left(\frac{\lambda_0}{\lambda_g}\right) \frac{\lambda_0^2}{ab} \left[A(\theta) \sin\theta + \frac{\lambda_g}{2a} B(\theta) \cos\theta\right]^2$$

where
$$A(\theta) = \frac{\cos(\pi\xi/2)}{1-\xi^2} + \frac{\cos(\pi\eta/2)}{1-\eta^2}$$

$$B(\theta) = \frac{\cos(\pi\xi/2)}{1-\xi^2} - \frac{\cos(\pi\eta/2)}{1-\eta^2}$$

[1] See references 15 and 18.

$$\xi = \frac{\lambda_0}{\lambda_g}\cos\theta - \frac{\lambda_0}{2a}\sin\theta$$

$$\eta = \frac{\lambda_0}{\lambda_g}\cos\theta + \frac{\lambda_0}{2a}\sin\theta$$

And for an inclined slot in the narrow face (Fig. 99e), the shunt conductance is

$$g = 0.131\left(\frac{\lambda_g}{\lambda_0}\right)\frac{\lambda_0^4}{a^3b}\left[\frac{\sin\theta\cos\left(\frac{\pi\lambda_0}{2\lambda_g}\sin\theta\right)}{1-\left(\frac{\lambda_0}{\lambda_g}\right)^2\sin^2\theta}\right]^2$$

Where mutual coupling exists between adjacent slots, the shunt admittance or series impedance presented by a slot will, in general, differ from the above values.[1]

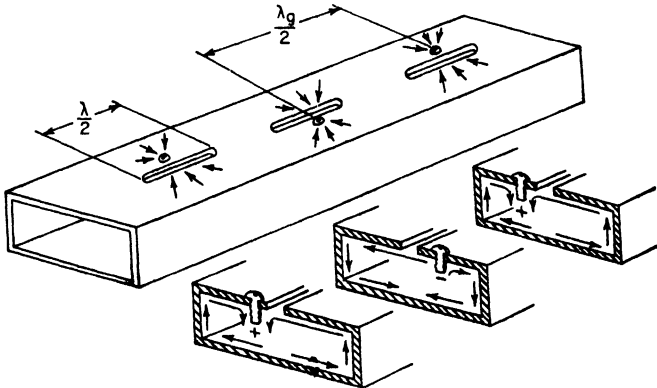



FIG. 100. Probe-fed slots on rectangular wave guide. The arrows show lines of current flow.

From a practical point of view dumbbell-shaped slots are used instead of rectangular slots. The perimeter of the slot, for resonance, is approximately equal to a wavelength. This not only decreases slot length but also simplifies machining. A thin sheet of dielectric placed over a slot will also decrease the resonant length.

Arrays of shunt slots are generally constructed by placing them $\lambda_g/2$ apart so they are effectively in parallel and by placing a short circuit at the far end $\lambda_g/4$ from the last slot so that the admittance in parallel with the last slot is zero. Then if there are k elements, the relative conductance of each slot is made equal to $1/k$. If the shunt elements are longitudinal slots cut into the broad wall of the wave-guide feed, the mutual coupling between slots is negligibly small.

The conductance of a longitudinal slot cut into the broad wall can be determined by measuring the input admittance of k slots in parallel because the mutual impedance is negligible. However, when mutual impedances between slots exist, the effective conductance of the slot (in the presence of the other elements) may be found by measuring the additional conductance produced when one slot is added to an array of identical slots. In practice, a number of slots (*e.g.*, 10) are cut and the input admittance determined. The input admittance is then again determined when additional slots are cut in sets of, say, 3. Eventually the total susceptance becomes constant and the conductance linearly proportional to k.

A nonradiating slot can be made to radiate by inserting a suitable probe into the wave guide adjacent to the slot, thereby introducing the necessary asymmetry in the current distribution for excitation. These radiating elements are called *probe-fed*

[1] See reference 344.

slots. The magnitude of the coupling can be controlled by adjusting probe depth, and the direction of field across the slot is determined by the side in which the probe is inserted. That is, the phase can be shifted 180 deg by switching the probe position. Examples are shown in Fig. 100.

Slots have a severe disadvantage at short wavelengths (1 cm), where they become so small that they have insufficient power-handling capacity and the tolerances on dimensions are impractical.

Broadside arrays of slot radiators are obtained by spacing the elements at intervals of $\lambda_g/2$ and changing the phase of successive elements by 180 deg. With shunt slots in the broad wall, the phase reversal is achieved by placing the slots on alternate sides of the center line of the wall, and for shunt-inclined slots on the narrow face, phase reversal is obtained by reversing the inclination of alternate slots. The phase of probe-fed slots is reversed by placing the probe on opposite sides on successive slots, as indicated above.

93. Surface-wave or Traveling-wave Antennas. A number of antennas can be classified as surface-wave or traveling-wave antennas. For example, dielectric-rod ("polyrod") antennas, dielectric-filled tapered-depth antennas, dielectric-sheet antennas, corrugated-surface antennas, helical-beam antennas (operating in the axial radiation mode), to name a few, are "end-fire" types of surface-wave antennas in which the wave velocity v along the antenna is usually less than the free-space velocity c. In this case, most of the energy transported by the wave is carried close to the antenna surface, but it will radiate efficiently if impeded by a discontinuity, such as a change of cross section. In addition to the above, the radiating or *"leaky"-wave-guide antenna* (also called the *channel-guide antenna* and *traveling-wave slot antenna*) and the *"serrated"-wave-guide antenna*[1] are another type of antenna that can be designed to radiate in directions other than in the end-fire direction (c/v less than 1) as well as in the end-fire direction. In general, this type of antenna can be easily scanned by changing the wave velocity along the aperture.

Most of these antennas, with the exception of the polyrod and helical antenna, are flush-mounted, that is, the radiating aperture is flush with the metallic surface in which they are mounted, and hence offer little aerodynamic drag. Most can be designed to give excellent pattern control by adjusting the amplitude and/or phase along the aperture. A variety of polarizations are available depending upon the antenna geometry and mode of excitation. A number of these antennas are filled with a dielectric material and those which are not can be designed with a dielectric sheet covering, which in either case acts as the antenna radome and permits pressurization of the whole system. Like the traveling-wave wire antenna, these antennas have a single feed point and are nonresonant, and so generally they have broad-band impedance characteristics.

FIG. 101. Cross section of the radiation pattern of a traveling-wave slot antenna.

The end-fire pattern is a pencil-type beam. For the flush-mounted antennas the end-fire pencil beam is necessarily modified by the presence of the metallic mounting surface. It appears that, in practice, maximum end-fire directivity is obtained when the wave velocity is such that the Hansen-Woodyard criterion[2] for increased directivity is satisfied, that is,

$$(\beta_s - \beta_0)l = \pi$$

where β_s = phase constant of the surface wave
β_0 = free-space phase constant
l = length of the antenna

[1] See references 354 and 374.
[2] See reference 47.

When radiation occurs in directions other than end-fire, the pattern has the shape of a hollow cone whose axis coincides with the long axis of the antenna, as illustrated in Fig. 101. In this case a good approximation to the radiation pattern of the traveling-wave antenna can be written as the product of the pattern of an element of the aperture and the pattern of a traveling wave which, if the rate of attenuation is small, is of the form $(\sin x)/x$, where x is a function of the polar angle θ, the antenna length, and the relative phase velocity of the traveling-wave field along the antenna. The effect of attenuation on the traveling-wave pattern is to broaden the main lobe, fill in the nulls between the side lobes (which is sometimes desirable), and raise the side-lobe level slightly. If the metallic surface is an infinite ground plane the pattern is independent of the ϕ coordinate; however, in practice it is modified by the size and shape of the surface.

94. Polyrod Antenna.[1] The dielectric-rod or polyrod antenna is a finite length of a tapered dielectric cylinder. If circular in cross section, the dominant mode is a

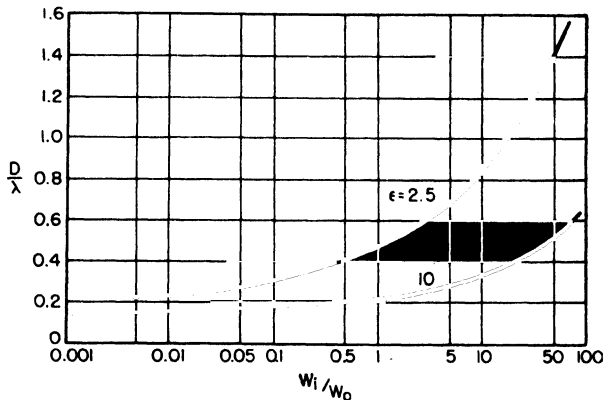



FIG. 102. Diameter in wavelengths as function of power ratio W_i/W_0, where W_i is the power inside the dielectric wave guide and W_0 is the power outside.

hybrid $TE_{1,1}$ mode, or, if rectangular, the dominant mode is a hybrid $TE_{1,0}$ mode. Neither has a cutoff frequency. Hence, this antenna is a broad-band radiator. However, at low frequencies most of the energy is outside the dielectric, *i.e.*, only loosely coupled to the guide. The ratio of the energy carried inside a circular rod to that outside is shown in Fig. 102 as a function of the diameter in wavelengths. When most of the energy is outside, the phase velocity v is approximately equal to the free-space velocity. When most of the energy is carried inside, the velocity approaches the value it would have if propagating in an unbounded region of the dielectric material. The relative phase velocities as a function of diameter for various dielectric constants ϵ are shown in Fig. 103. The transition region is quite narrow for the higher dielectric constants. If the rod is gradually tapered from a diameter at which most of the energy is inside the dielectric to a diameter at which most of the energy is outside, the effective area of the wave front is large, and this wave front, acting as a secondary source of radiation, produces a pattern similar to a horn. Once most of the energy in the traveling wave is outside the dielectric, the wave velocity is approximately the free-space velocity and the dielectric can be terminated without disturbing the wave.

Because of its properties the dielectric-rod antenna is a broad-band radiator. An example of a practical antenna is shown in Fig. 104.[2]

[1] See reference 13.
[2] See references 9 and 363.

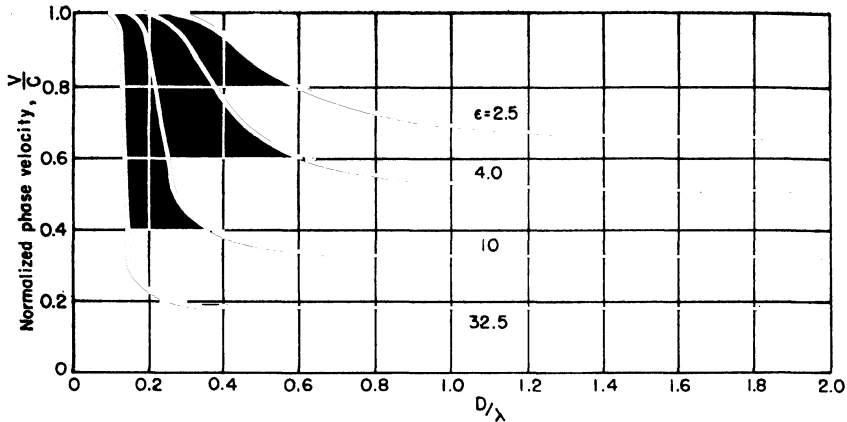



FIG. 103. Ratio of phase velocity along the dielectric wave guide to the velocity of light as a function of diameter in wavelengths.

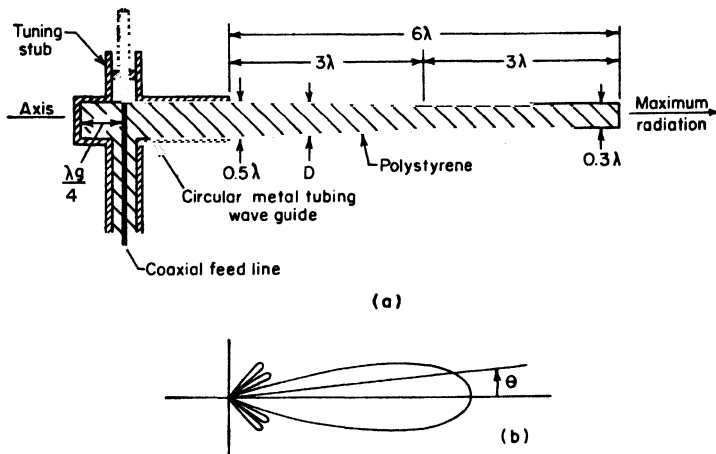



FIG. 104. A dielectric-rod antenna.

The directivity of the dielectric-rod antenna of length L is given approximately by

$$g \doteq \frac{8L}{\lambda_0}$$

The half-power beam width is given approximately by

$$\text{Half-power BW} \doteq \frac{60}{\sqrt{L/\lambda_0}} \quad \text{deg}$$

where L/λ_0 = antenna length, free-space wavelengths.

The diameter of the circular wave-guide feed must be greater than $0.58\lambda_0/\sqrt{\epsilon_r}$, where ϵ_r is the relative dielectric constant of the dielectric material filling the guide.

95. Dielectric-filled Tapered-depth Antenna.[1] This antenna is equivalent to a flush-mounted polyrod. Basically, it consists of a dielectric-filled wave guide excited

[1] See reference 375.

in the dominant $TE_{1,0}$ mode, as shown in Fig. 105. The cross section of the "radiating" portion of the antenna is tapered in thickness. Typical principal plane patterns for a tapered-depth antenna (with uniform taper) mounted in a finite ground plane are also shown in Fig. 105. In this figure the H plane is the plane of the metallic ground plane; the E plane is perpendicular to the ground plane.

The shape of the E-plane pattern is influenced by the shape and length of the metallic surface beyond the end of the antenna and by radiation from spurious modes

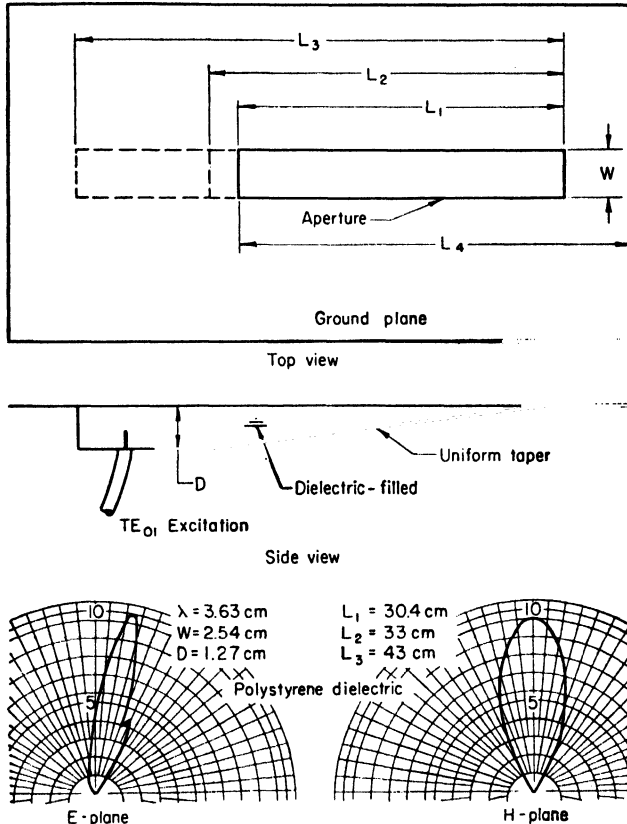



FIG. 105. Tapered-depth end-fire antenna and typical radiation pattern.

set up at the feed end of the aperture due to improper mode conversion at this point. Some pattern control can be obtained by the taper of the dielectric.[1] Very practical antennas, however, have been obtained with uniform tapers. The shape of the H-plane pattern is determined by the aperture width W.

The wave velocity in this antenna is a function of the depth, width, and dielectric constant of the material. The relative velocity ratio c/v for polystyrene dielectric ($\epsilon_r = 2.5$) is shown in Fig. 106 as a function of depth and width.[2] As is true of the dielectric-rod antenna, the wave velocity can vary between the limits $\sqrt{\epsilon_r}$ and 1.0 over a wide frequency range. Hence, the tapered-depth antenna results in an end-fire antenna whose practical band width is limited only by cutoff in the feed section (which

[1] See reference 332.
[2] See reference 375.

can be extended by using a ridged wave guide) and the discontinuity at the beginning of the aperture.

Large-aperture wide-band tapered-depth antennas can be constructed simply as shown in Fig. 107. The flared antenna is effectively a flush-mounted horn. The rectangular aperture is excited by an H-plane sectoral horn with a dielectric lens to correct the phase front. The lens can be constructed as part of the dielectric slab.

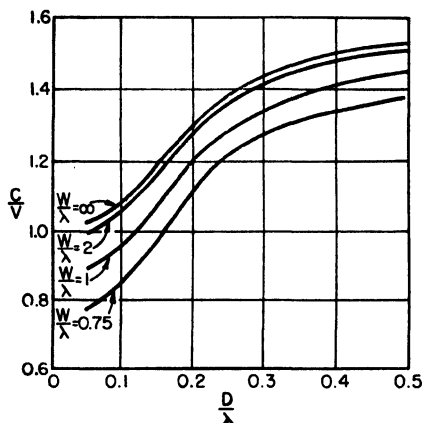



FIG. 106. Calculated c/v as a function of wave-guide geometry for polystyrene dielectric.

This principle has been applied in the construction of a tapered circular dielectric-disk antenna to produce a high-gain omnidirectional pattern in the plane of the disk.[1] It consists of a circular dielectric disk, tapered uniformly in thickness from the center to the edge, mounted on a metallic ground plane. The disk is excited from beneath by a dipole—the extended center conductor of a coaxial line. Only the lowest-order (dominant) surface wave mode can exist in the disk as long as its thickness is not greater than $\lambda_0/4\sqrt{\epsilon_r - 1}$, where ϵ_r is the relative dielectric constant. It has been found experimentally that the uniform taper results in the highest gain and lowest side-lobe level.

96. Corrugated-surface Antenna. A corrugated surface similar to the one shown in Fig. 108 will act as a guide for surface waves, with the direction of propagation perpendicular to the length of the corrugations, provided the loading due to the corrugations (slots) is series-inductive.[2] The mode is transverse magnetic (TM), since there is no component of H in the direction of propagation. This occurs when the

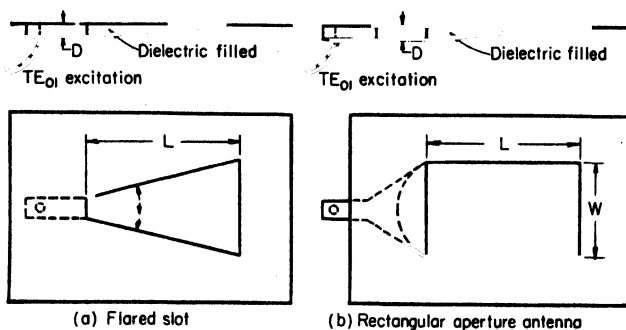



FIG. 107. Large-aperture tapered-depth antenna.

depth of slots is less than $\lambda/4$ (plus an integral number of half wavelengths). The phase velocity along the surface is a function of slot depth, varying from free-space velocity at zero depth to zero velocity at a depth of $\lambda/4$. The phase velocity is also a function of surface width w; however, it has been found that the phase velocity is approximately that of a wave on a corrugated surface infinitely wide when w is greater

[1] See reference 352.
[2] See reference 371.

than about 3λ. The attenuation on these surfaces is due chiefly to metal losses and varies between 0.2 and 0.5 db per m.

A corrugated surface can be made to radiate either by matching the impedance of the wave on this surface to free space by means of tapered corrugated sections or by introducing abrupt periodic discontinuities in the surface. Little radiation occurs from the edges of such an antenna. An example is shown in Figs. 109 and 110.[1] The

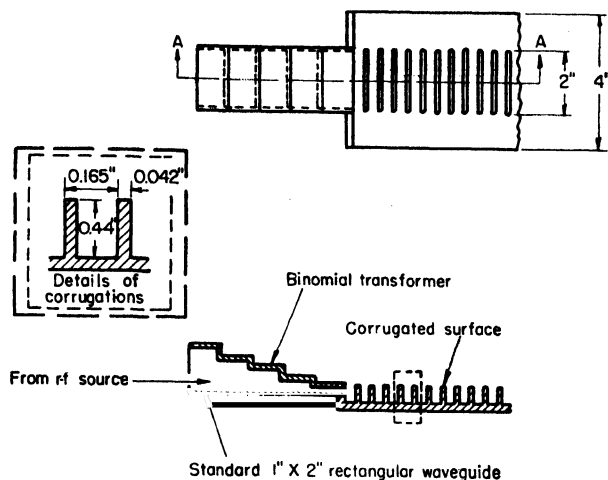



FIG. 108. Single flat corrugated surface with wave-guide feed.

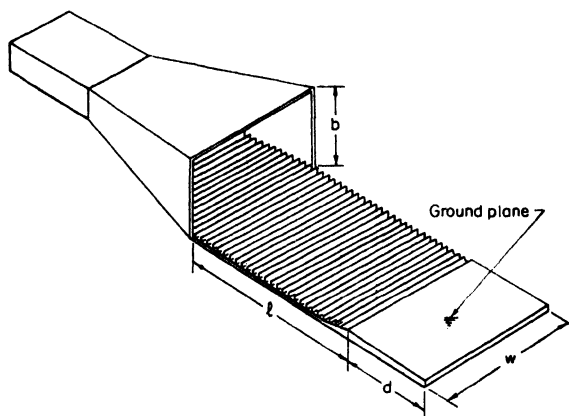



FIG. 109. Corrugated-surface antenna.

antenna consists of three main parts: the wave-guide feed, generally propagating the dominant $TE_{1,0}$ mode, a *mode transformer* (or *horn launcher*), and the corrugated surface. If the side and top walls of the launcher are gradually flared out over several wavelengths before they are terminated and a tapered corrugated section is inserted in the launcher, a surface wave will be efficiently launched on the corrugated surface with a good match.[2] In general, the radiation pattern consists of a superposition of

[1] See reference 345.
[2] See reference 343.

the radiation from the mouth of the launcher, the corrugated section, and the currents flowing on the finite ground plane beyond the corrugated section. If the contribution of the launcher is to be kept small, its mouth must also be small. This is contrary to the requirements for a good match; hence, the optimum height b (since w will be determined by pattern requirements) should be determined by experiment. b is a

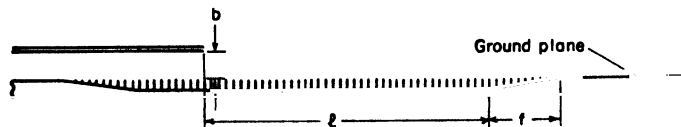



FIG. 110. Simplified antenna system.

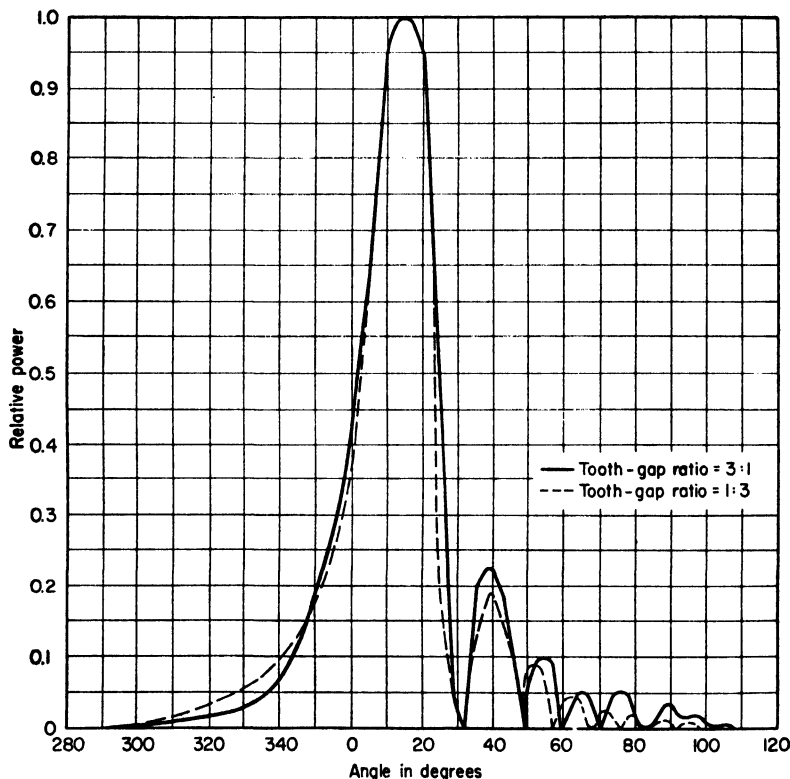



FIG. 111. Experimental radiation patterns for two corrugated surfaces 7.33λ long, one having a tooth-to-gap ratio of 3:1 and the other a tooth-to-gap ratio of 1:3, each terminated by a λ/2 ground plane.

function of both β_s and l. It might be possible to adjust the amplitude and phase of the radiation from the launcher to improve the radiation pattern.

If h is the depth of the corrugation, G the gap width, and T the tooth width, the surface dimensions that will permit propagation must satisfy the relation

$$\cot \beta_0 h = \frac{G}{G + T} \frac{1}{\sqrt{(\beta_s/\beta_0)^2 - 1}}$$

where $\beta_0 = 2\pi/\lambda_0$ = free-space phase constant
$\beta_s = 2\pi/\lambda_s$ = phase constant of the surface wave
$\beta_s/\beta_0 = c/v$

This relation holds if the number of corrugations per wavelength is large (greater than 19 per wavelength) and the tooth width/gap width ratio is small. The latter condition holds strictly when the corrugated surface is being used as a transmission line because it has been found in practice that there was no significant difference in the radiation patterns for surfaces with tooth width/gap width ratios of 3:1, 1:1, and 1:3. An example of the type of radiation pattern (E plane) obtained from the corrugated-surface antenna and the effect of tooth width is shown in Fig. 111 ($\theta = 0$ deg corresponds to the direction of the long axis of the surface in the direction of propagation). The value of $\beta_s/\beta_0 = c/v$ in this relation is generally chosen to satisfy the Hansen-Woodyard relation for optimum end-fire directivity, *i.e.*, $(\beta_s - \beta_0)l = \pi$.

The width of the surface is determined by the required H-plane (horizontal) beam width. The beam width and tilt angle (angle above the ground plane) of the main beam are functions of the length of the corrugated section and the length and shape of the ground plane beyond it.

To effect a match to free space from the end of the corrugated surface, a tapered section f a few wavelengths long is inserted between the end of the corrugated section and the ground plane.

97. Leaky-wave-guide or Traveling-wave-slot Antenna.[1] The leaky-wave-guide or traveling-wave-slot antenna is similar to the traveling-wave wire antenna, but it is far more versatile because the phase velocity and the rate of radiation of the fields in the antenna, and hence its pattern, can be controlled by the geometry of the antenna. (This type differs from those previously mentioned primarily in that $c/v \leq 1$.) In general, this antenna consists of either a long slot cut so that it is parallel to the long axis of the guide or an array of closely spaced nonresonant slots cut into one wall of a wave guide. The guided waves are characterized by a complex propagation constant since energy in the waves escapes from the wave guide.

There are four typical field configurations associated with traveling-wave-slot antennas which are excited by a uniform wave guide whose axis is parallel to the length of the slot. These are: (1) the conventional transverse electric (no tangential E parallel to slot length); (2) transverse magnetic (no tangential H parallel to slot length); (3) a hybrid with a negligible transverse E component; and (4) a hybrid with a negligible normal H component.

Any of these configurations can be excited by proper positioning of a long slot in the wall of a uniform wave guide of rectangular or circular cross section. To obtain the transverse electric excitation a narrow slot must be placed where the current in the wall of the wave guide is perpendicular to the slot. The transverse magnetic excitation is obtained from an air-filled wave guide having a wide slot (about $\lambda/2$) positioned so that the current flow is parallel to the slot length. The hybrid excitation (3) is obtained under the same conditions as the TM excitation, except that the wave guide is filled with dielectric having a dielectric constant of 2.0 or greater. The hybrid excitation (4) is obtained from a wave guide partly filled with dielectric (such as polystyrene) which is excited in the lowest-order (hybrid) wave-guide mode. (This mode also exists in the tapered-depth antenna.)

The propagation constants for TE and TM types of traveling-wave-slot antennas are shown in Figs. 112 to 115 in the form of the velocity ratio c/v and attenuation constant α in nepers per wavelength, where c is the free-space velocity and v is the phase velocity along the slot aperture.[2]

The radiation pattern has the shape of a hollow cone with its axis coincident with the slot axis (see Fig. 101). For an antenna with a uniform cross section (*i.e.*, c/v is a constant, and the strength of the aperture field decays exponentially), the radiation pattern, assuming the aperture is in an infinite ground plane, is given by

$$|E(\theta,\phi)| = k|f(\theta,\phi)| \left\{ \frac{1 - 2e^{-\alpha l} \cos [\beta_0 l(c/v - \cos\theta)] + e^{-2\alpha l}}{(c/v - \cos\theta)^2 + (\alpha/\beta_0)^2} \right\}^{1/2}$$

[1] See references 240, 348, 349, and 370.
[2] See reference 349.

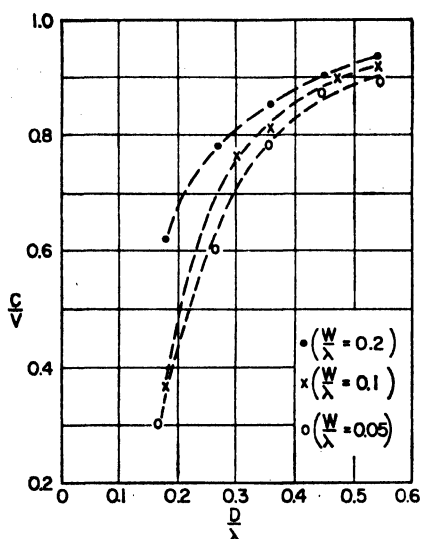



FIG. 112. Measured velocity ratios c/v of an air-filled *TE* traveling-wave-slot antenna.

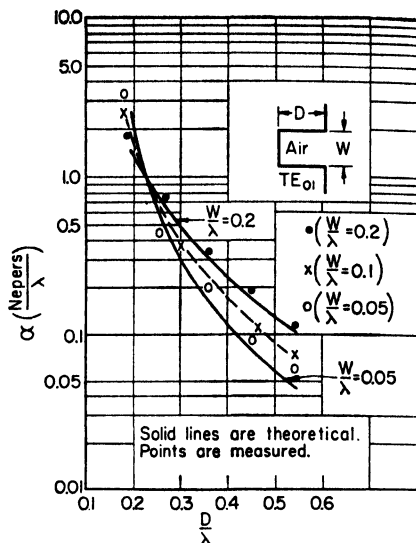



FIG. 113. Attenuation constants of an air-filled *TE* traveling-wave-slot antenna.

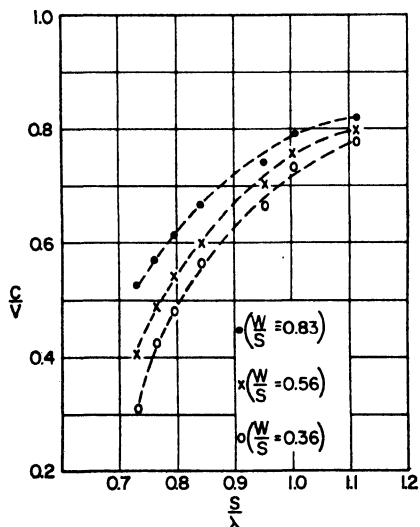



FIG. 114. Measured velocity ratios of an air-filled *TM* traveling-wave-slot antenna of square cross section with $TM_{1,1}$ excitation.

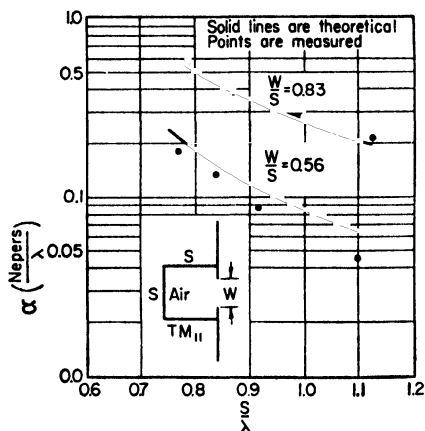



FIG. 115. Attenuation constants of an air-filled *TM* traveling-wave-slot antenna of square cross section with $TM_{1,1}$ excitation.

where $f(\theta,\phi)$ = the pattern of a differential element of the aperture (= $\sin\theta$ for *TE* types, for example) .

α = attenuation constant, nepers per m

β_0 = free-space phase constant, radians/m

l = length of aperture, m

k = constant

If α is small, then to a good approximation the direction of maximum radiation θ_m is given by

$$\theta_m = \cos^{-1} (c/v)$$

This also neglects the influence of the pattern of the differential element, but if l is large ($>5\lambda$), this is still a good approximation.

Beam shaping can be accomplished with this type of antenna by using a slot whose width W varies with distance along the slot to give the desired amplitude distribution. That is, the rate of attenuation (radiation) along the antenna can be controlled by the aperture width. If this variation in W also results in an undesired change in the phase velocity, the depth or cross section can be modified to compensate for it.

An approximation to the desired variation in the attenuation constant along the slot to obtain an amplitude distribution represented by the function A is

$$\alpha(z) = \frac{1}{2} \frac{A^2}{\int_z^l A^2 \, dz + \frac{P(l)}{P(0) - P(l)} \int_0^l A^2 \, dz}$$

where $P(0)$ = input power at $z = 0$
$P(l)$ = power left over at $z = l$
l = length of slot aperture

It has been found from experience that the exact array-design procedure (described in Sec. 29) must be employed in designing arrays of traveling-wave antennas when the element spacing is less than about 2λ.

The mutual impedance between traveling-wave-slot antennas may be analyzed by picturing the traveling-wave slot as a directional coupler. When two such slots are placed along parallel lines, the field induced in one due to energization of the other will consist primarily of a wave traveling in the same direction as the wave in the slot which is energized directly. This wave will be reflected at the end of the parasitic slot and will travel back to its input terminals. However, it is attenuated by radiation as it travels along the slot so that the amount available at the input terminals of the parasitic slot is very small. If the slots are long enough to be good directional antennas, the coupling between adjacent input terminals is very small. In practice it has proved to be negligible.

The magnitude of the mutual impedance is quite different for the *TE*, *TM*, or hybrid slots. For example, the equivalent magnetic dipoles associated with parallel slots are broadside to broadside for *TE* operation, whereas they are end to end for hybrid operation (3) (no transverse *E*). Consequently the mutual-impedance effects associated with hybrid operation (3) are much less than with *TE* operation.

References

General Textbooks:

1. Aharoni, J.: "Antennae; An Introduction to Their Theory," Oxford, 1946.
2. Atwood, S. S.: "Electric and Magnetic Fields," Wiley, 1941.
3. American Radio Relay League: "ARRL Antenna Handbook," 1955.
4. Brainerd, J. G. (editor): "Ultra-high Frequency Techniques," Van Nostrand, 1942.
5. Bronwell, A. B., and R. E. Beam: "Theory and Application of Microwaves," McGraw-Hill, 1947.
6. Fry, D. W., and F. K. Howard: "Aerials for Centimeter Wavelength," Cambridge, 1950.
7. Jordan, E. C.: "Electromagnetic Waves and Radiating Systems," Prentice-Hall, 1950.
8. King, R. W. P., H. R. Mimno, and A. H. Wing: "Transmission Lines, Antennas, and Wave Guides," McGraw-Hill, 1945.
9. Kraus, J. D.: "Antennas," McGraw-Hill, 1950.
10. Laport, E. A.: "Radio Antenna Engineering," McGraw-Hill, 1952.
11. Ramo, S., and J. R. Whinnery: "Fields and Waves in Modern Radio," Wiley, 1944.
12. Schelkunoff, S. A.: "Electromagnetic Waves," Van Nostrand, 1943.
13. ——— and H. T. Friis: "Antenna Theory and Practice," Wiley, 1952.
14. ———: "Advanced Antenna Theory," Wiley, 1952.
15. Silver, S.: "Microwave Antenna Theory and Design," McGraw-Hill, 1949.
16. Stratton, J. A.: "Electromagnetic Theory," McGraw-Hill, 1941.

17. Terman, F. E.: "Radio Engineers' Handbook," McGraw-Hill, 1943.
18. Watson, W. H.: "Transmission and Antenna Systems," Oxford, 1947.

Radiation and Antenna Theory:

19. Ballantine, Stuart: On the Radiation Resistance of a Simple Vertical Antenna at Wave Lengths below the Fundamental, *Proc. IRE*, **12,** 823, December, 1924.
20. ———: On the Optimum Transmitting Wave Length for a Vertical Antenna over Perfect Earth, *Proc. IRE*, **12,** 833, December, 1924.
21. Barrow, W. L.: Impedance of a Vertical Half-wave Antenna above an Earth of Finite Conductivity, *Proc. IRE*, **23,** 150, February, 1935.
22. Bashenoff, V. I., and N. A. Mjasoedoff: Effective Height of Closed Aerials, *Proc. IRE*, **19,** 984, June, 1931.
23. Bechmann, R.: Calculation of Electric and Magnetic Field Strengths of Any Oscillating Straight Conductors, *Proc. IRE*, **19,** 461, March, 1931.
24. ———: Calculation of Radiation Resistance of Antennas and Antenna Combinations, *Proc. IRE*, **19,** 471, August, 1931.
25. Cox, C. R.: Mutual Impedance between Vertical Antennas of Unequal Heights, *Proc. IRE*, **35,** 1367, November, 1947.
26. Born, H.: Energy Distribution in the Near Field of Electro-magnetic Radiators, in Particular in Front of Diaphragms and Reflectors (treated by acoustical methods); abstract, *Wireless Engr.*, **21,** 188, April, 1944.
27. Bouwkamp, C. J.: Hallens Theory for a Straight, Perfectly Conducting Wire, Used as a Transmitting or Receiving Aerial, *Physica*, **9,** 609–631, July, 1942.
28. Brown, G. H., and R. King: High-frequency Models in Antenna Investigations, *Proc. IRE*, **22,** 457, April, 1934.
29. Brown, G. H.: Phase and Magnitude of Earth Currents near Radio Transmitting Antennas, *Proc. IRE*, **23,** 168, February, 1935.
30. ——— and O. M. Woodward: Experimentally Determined Impedance Characteristics of Cylindrical Antennas, *Proc. IRE*, April, 1945, pp. 257–262.
31. Burgess, R. E.: Aerial Characteristics; Relation between Transmission and Reception, *Wireless Engr.*, **21,** 154, April, 1944.
32. Carter, P. S.: Circuit Relations in Radiating Systems and Applications to Antenna Problems, *Proc. IRE*, **20,** 1004, June, 1932.
33. Colebrook, F. M.: Experimental and Analytical Investigation of Earthed Receiving Aerials, *J. IRE*, **71,** 235, July, 1932.
34. ———: Electric and Magnetic Fields for a Linear Radiator Carrying a Progressive Wave, *J. IEE*, February, 1940, p. 169.
35. ——— and A. C. Gordon-Smith: Method of Calibrating Field Strength Measuring Set, *J. IEE*, March, 1941, p. 15.
36. Cronshey, R. W.: Calculating Antenna Radiation Patterns, *Elec. Eng.*, **63,** 331, September, 1944.
37. Cutting, Fulton: Simple Method of Calculating Radiation Resistance, *Proc. IRE*, **10,** 129, April, 1922.
38. Forbes, Henry C.: Re-radiation from Tuned Antenna Systems, *Proc. IRE*, **13,** 363, June, 1925.
39. Foster, Ronald M.: Directive Diagrams of Antenna Arrays, *Bell System Tech. J.*, **5,** 292, January, 1926.
40. Friis, H. T.: A Note on a Simple Transmission Formula, *Proc. IRE*, **34,** 254, May, 1946. (Effective areas of transmitting and receiving antennas.)
41. Giacoletto, L. J.: Antenna Radiation Chart, *Electronics*, **12,** 35, July, 1939.
42. Gray, M. C.: Modification of Hallen's Solution of the Antenna Problem, *J. Applied Phys.*, **15,** 61, January, 1944.
43. Green, E.: Extended Aerial Systems; Calculating the Polar Diagrams, *Wireless Eng.*, **19,** 195, May, 1942.
44. Grosskopf, J.: Radiation Field of a Vertical Transmitting Dipole over Stratified Ground, *Wireless Engr.*, **20,** 245, May, 1943.
45. Grover, F. W.: Methods, Formulas, and Tables for the Calculation of Antenna Capacity, *Nat. Bur. Standards, Sci. Papers* 568–569, 1928.
46. Hansen, W. W., and J. G. Beckerley: Concerning New Methods of Calculating Radiation Resistance, Either with or without Ground, *Proc. IRE*, **24,** 1594, December, 1936.
47. ——— and J. R. Woodyard: A New Principle in Directional Antenna Design, *Proc. IRE*, March, 1938, p. 333.
48. Harrison, C. W., Jr., and R. King: Receiving Antenna in a Plane-polarized Field of Arbitrary Orientation, *Proc. IRE*, **32,** 35, January, 1941.
49. ———: Mutual Impedance of Antennas, *J. Applied Phys.*, **14,** 306, June, 1943.
50. ———: Approximate Representation of the Electromagnetic Field in the Vicinity of a Symmetrical Radiator, *J. Applied Phys.*, **15,** 544, July, 1944.
51. Hines, J. N., V. H. Rumsey, and T. E. Tice: On the Design of Arrays, *Proc. IRE*, **42,** 1262, August, 1954.
52. Howe, G. W. O.: Equivalent Inductance and Capacity of an Aerial with Inserted Tuning Coil or Condenser, *Exptl. Wireless*, **5,** 297, June, 1928.
53. ———: Applying Transmission Line Theory to Aerials, *Wireless Engr.*, January, 1939, p. 1.
54. ———: Calculation of Aerial Capacitance, *Wireless Engr.*, **20,** 157, April, 1943.
55. ———: Radiation Resistance of a Half-wave Dipole Aerial, *Wireless Engr.*, **22,** 153, April, 1945.

56. Jachnow, W.: Mutual Impedance of Inclined Rectilinear Conductors with Progressive Waves, *ENT*, July, 1939, p. 177.
57. ———: Theoretical Investigations of Radiation Diagrams and Radiation Resistance for Progressive Waves of Various Phase Velocities, abstract, *Wireless Engr.*, **20**, 87, February, 1943.
58. Jordan, E. C.: Acoustic Models of Radio Antennas, *Ohio State Univ. Eng. Theories*, **10** (3), May, 1941.
59. ——— and W. L. Everitt: Acoustic Models of Radio Antennas, *Proc. IRE*, April, 1941, p. 186.
60. Kelvin, Wm.: Radiation Field of an Unbalanced Dipole, *Proc. IRE*, **34**, 440, July, 1946.
61. King, R.: Approximate Representation of the Distant Field of Linear Radiators, *Proc. IRE*, **29**, 458, August, 1941.
62. ——— and F. C. Blake, Jr.: Self-impedance of a Symmetrical Antenna, *Proc. IRE*, **30**, 335, July, 1942.
63. ——— and C. W. Harrison, Jr.: Distribution of Current along a Symmetrical Center-driven Antenna, *Proc. IRE*, **31**, 548, October, 1943.
64. ———: Coupled Antennas and Transmission Lines, *Proc. IRE*, **31**, 626, November, 1943.
65. ——— and C. W. Harrison, Jr.: Receiving Antenna, *Proc. IRE*, **32**, 18, January, 1944.
66. ——— and ———: Impedance of Short, Long, and Capacitively Loaded Antennas with a Critical Discussion of the Antenna Problem, *J. Applied Phys.*, **15**, 170, February, 1944.
67. ——— and ———: Mutual and Self-impedance for Coupled Antennas, *J. Applied Phys.*, **15**, 481, June, 1944.
68. Labus, J.: Recherische Ermittlung der Impedanz von Antennan, *Hochfrequenztechn.*, January, 1933, p. 17. (Mutual impedance theory.)
69. Laport, E. A.: Calculation of Vertical Antenna Radiation Patterns, *Electronics*, February, 1935.
70. Levin, S. A., and C. J. Young: Field Distribution and Radiation Resistance of a Straight Vertical Unloaded Antenna Radiating at One of Its Harmonics, *Proc. IRE*, **14**, 675, October, 1926.
71. McPetrie, J. S.: Graphical Method for Determining the Magnitude and Phase of the Electric Field in the Neighborhood of an Antenna Carrying a Known Distribution of Current, *J. IEE*, **69**, 290, February, 1931.
72. ———: Method for Determining the Effect of the Earth on the Radiation from Aerial Systems, *J. IEE*, **70**, 382, March, 1932.
73. ——— and J. A. Saxon: Theory and Confirmation of Calibration of Field Strength Sets by Radiation, *J. IEE*, May, 1941, p. 11.
74. Moullin, E. B.: Radiation Resistance of Aerials Whose Length Is Comparable with the Wavelength, *J. IEE*, **78**, 540–563, May, 1936.
75. ———: Radiation Resistance of Surfaces of Revolution, Such as Cylinders, Spheres, and Cones, *J. IEE*, **88**, 50, March, 1941.
76. Murray, F. H.: Mutual Impedance of Two Skew Antenna Wires, *Proc. IRE*, **21**, 154, January, 1933.
77. Neiman, M. S.: Principle of Reciprocity in Antenna Theory, *Proc. IRE*, **31**, 666, December, 1943.
78. Neissen, K. F.: Ground Absorption for Horizontal Dipole Aerials, *Ann. Physik*, July, 1938, p. 444.
79. ———: Choice between Horizontal and Vertical Dipole Aerials for Minimum Earth Absorption with Given Wavelength and Soil Type, *Ann. Physik*, November, 1938, p. 403.
80. ———: Radiation from a Dipole, *Ann. Physik*, February, 1937, p. 209.
81. ——— and G. deVries: Receiving Impedance of a Receiving Antenna, *Physica*, July, 1939, p. 601.
82. ———: Calculation of Field Strength of a Half-wave Aerial, *Physica*, September, 1940, p. 23.
83. Pistolkors, A. A.: Radiation Resistance of Beam Antennas, *Proc. IRE*, **17**, 562, March, 1929.
84. Ramsey, R. R.: Radiation and Induction, *Proc. IRE*, **21**, 1586, November, 1933.
85. Ratcliffe, J. A., L. G. Vedy, and A. F. Wilkins: Spreading of Electromagnetic Waves from a Hertzian Dipole, *J. IEE*, **70**, 522, May, 1932.
86. Schelkunoff, S. A.: On Diffraction and Radiation of E-m Waves, *Phys. Rev.*, August, 1939, p. 308.
87. ———: General Radiation Formula, *Proc. IRE*, October, 1939, p. 660.
88. ———: Theory of Antennas of Arbitrary Size and Shape, *Proc. IRE*, September, 1941, p. 493.
89. ——— and C. B. Feldman: Radiation from Antennas, *Proc. IRE*, **30**, 511, November, 1942.
90. ———: Mathematical Theory of Linear Arrays, *Bell System Tech. J.*, **22**, 80, January, 1943.
91. ———: Antenna Theory and Experiment, *J. Applied Phys.*, **15**, 54, January, 1944.
92. Siegel, E. M.: Wavelength of Oscillations along Transmission Lines and Antennas, *Texas Univ. Eng. Research Ser.* 32, 1, 1940.
93. Southworth, G. C.: Certain Factors Affecting the Gain of Directive Antennas, *Proc. IRE*, **18**, 1502, September, 1930.
94. Sommerfeld, A., and F. Renner: Radiation Energy and Earth Absorption for Dipole Antennae, *Wireless Engr.*, **19**, 351, 409, 457, August–October, 1942.
95. Stratton, J. A., and L. J. Chu: Diffraction Theory of E-m waves, *Phys. Rev.*, July, 1939, p. 99.
96. Thomson, W. T.: Development of the General Antenna Array Equation, *J. Applied Phys.*, **15**, 420, May, 1944.
97. Walmsley, T.: Impedance Characteristics of Short-wave Dipoles, *Phil. Mag.*, June, 1938, p. 981.
98. Wells, N.: Aerial Characteristics, *J. IEE*, **89**, Pt. 3, 76, June, 1942.
99. Wheeler, H. A.: Fundamental Limitation of Small Antennas, *Proc. IRE*, **35**, 1479, December, 1947.
100. Wilmotte, R. M.: Distribution of Current in a Transmitting Antenna, *J. IEE*, **66**, 617, June, 1928.
101. Wise, W. Howard: Grounded Condenser Antenna Radiation Formula, *Proc. IRE*, **19**, 1684, September, 1931.
102. Zinke, C.: Fundamental Consideration of the Current and Potential Distribution on Antennas, *Wireless Engr.*, September, 1941, p. 377.

L-f Antennas:

103. Alexanderson, E. F. W.: Trans-Atlantic Radio Communication, *Proc. IRE*, August, 1920.
104. ———: Transoceanic Radio Communication, *Gen. Elec.*, *Rev.* October, 1920.
105. Ashbridge, N., H. Bishop, and B. N. MacLarty: Droitwich Broadcasting Station, *J. IEE*, **77**, 447, 460; Discussion, 474, October, 1935.
106. Bailey, A., S. W. Dean, and W. T. Wintringham: Receiving System for Longwave Transatlantic Radiotelephony, *Bell System Tech. J.*, **8**, 309, April, 1929.
107. Beverage, H. H., C. W. Rice, and E. W. Kellogg: The Wave Antenna, *Trans. AIEE*, **42**, 215, 1923.
108. Eckersley, T. L.: An Investigation of Transmitting Aerial Resistances, *J. IEE*, **60**, 581–605, 1922. (Studies of direct earth and counterpoise capacity grounds.)
109. Kear, F. G.: Phase Synchronization in Directive Antenna Arrays with Particular Application to the Radio Range Beacon, *J. Research Nat. Bur. Standards*, **11**, 123, July, 1933.
110. Lindenblad, N., and W. W. Brown: Main Considerations in Antenna Design, *Proc. IRE*, **14**, 291, June, 1926.
111. Mann, H. F., and F. Hollinghurst: Replacement of Main Aerial System at Rugby Station, *POEE J.*, April, 1940, p. 22.
112. Morgan, M. G.: Increasing Radiation at Low Frequencies, *Electronics*, July, 1940, p. 33.
113. Rosseler, G., and K. Vogt: Investigations on Umbrella Aerials, abstract, *Wireless Engr.*, **20**, 141, March, 1943.
114. Watson-Watt, R. A., and J. F. Herd: An Instantaneous Direction Reading, Radio Goniometer, *J. IEE*, **64**, 611, 1926.

M-f Antennas:

115. Abbott, F. R.: Design of Optimum Buried-conductor R.F. Ground System, *Proc. IRE*, **40**, 846, July, 1952.
116. Andrew, V. J., and A. M. McGregor: Low Cost Wooden Mast; Station WJBC, *Electronics*, **9**, 45, May, 1936.
117. Boudoux, P.: Current Distribution and Radiation Properties of the Shunt-excited Antenna, *Proc. IRE*, June, 1940, p. 271.
118. Brown, G. H.: Critical Study of the Characteristics of Broadcast Antennas as Affected by Antenna Current Distribution, *Proc. IRE*, **24**, 48, January, 1936.
119. ———: Directional Antennas, *Proc. IRE*, **25**, 78, January, 1937.
120. ——— and J. G. Leitch: Fading Characteristics of the Top-loaded WCAU Antenna, *Proc. IRE*, **25**, 583, May, 1937.
121. ———, R. F. Lewis, and J. Epstein: Ground System as a Factor in Antenna Efficiency, *Proc. IRE*, **25**, 753, June, 1937.
122. ———: A Radical Ground System Chart, *Electronics*, January, 1938, p. 33.
123. ———: WTAR Directional Array, *Electronics*, January, 1938, p. 38.
124. ———: Consideration of the Radio Frequency Voltages Encountered by the Insulating Materials of Broadcast Towers, *Proc. IRE*, September, 1939, p. 566.
125. ——— and J. M. Baldwin: Adjusting Unequal Tower Broadcast Arrays, *Electronics*, **16**, 118, December, 1943.
126. Chamberlain, A. B., and W. B. Lodge: Broadcast Antenna, *Proc. IRE*, **24**, 11, January, 1936.
127. Cox, C. R.: Phasing Networks for Broadcast Arrays; Graphical Methods Applied, *Electronics*, **17**, 120, June, 1944.
128. Dewitt, J. H., Jr., and A. D. Ring: Significant Radiation from Directional Antennas of Broadcast Stations for Determining Sky-wave Interference at Short Distances, *Proc. IRE*, **32**, 668, November, 1944.
129. Doherty, W. H.: Operation of AM Broadcast Transmitters into Sharply Tuned Antenna Systems, *Proc. IRE*, **37**, 729, July, 1949.
130. Duttera, W. S.: Directional Antenna Chart, *Electronics*, February, 1940, p. 33.
131. ———: Design Chart for Phase Shifting and Amplitude Control Networks, *Electronics*, **15**, 53, October, 1942.
132. ———: Adjustment of Directional Antennas, *Electronics*, **16**, 91, April, 1943.
133. Eckersley, P. P., T. L. Eckersley, and H. L. Kirke: Design of Transmitting Aerials for Broadcasting Stations, *J. IEE*, **67**, 507, April, 1929.
134. Eppen, F., and A. Gothe: Ueber die schwundvermindernde Antenne des Rundfunksenders Breslau, *ENT*, **10** (4), 1933. (Top-loaded vertical wire on wood tower for antifading.)
135. Everest, F. A., and W. S. Pritchard: Horizontal Polar Pattern Calculator for Directional Broadcast Antennas, *Proc. IRE*, June, 1941, p. 355.
136. FCC: Part 3, Radio Broadcast Services, January, 1956.
137. Feld, J.: Radio Antenna Suspended from 1,000 Foot Towers, *J. Franklin Inst.*, **239**, 363, May, 1945.
138. Fitch, W. A., and W. S. Duttera: Measurement of Broadcast Coverage and Antenna Performance, *RCA Rev.*, July, 1938, p. 396.
139. Gihring, H. E., and G. H. Brown: General Consideration of Tower Antennas for Broadcast Use, *Proc. IRE*, **23**, 311, April, 1935.
140. Glas, E. T.: Efficiency-rating of Transmitting Aerials for Broadcasting Distribution, *Exptl. Wireless*, **7**, 665, December, 1930.
141. Harbich, H., and W. Hahnemann: Wirksame Bekampfung des Nahschwundes im Rundfunk durch Dendeantennengebilde bestimmter Form, *ENT*, **9** (10), October, 1932. (Treatment of circular arrays for antifading.)

142. Harmon, R. N.: Some Comments on Broadcast Antennas, *Proc. IRE*, **24**, 36, January, 1936.
143. Hutton, W. G., and R. M. Pierce: Mechanical Calculator for Directional Antenna Patterns, *Proc. IRE*, **30**, 233, May, 1942.
144. Jeffers, C. L.: An Antenna for Controlling the Nonfading Range of Broadcasting Stations, *Proc. IRE*, **36**, 1426, November, 1948.
145. Kear, F. G.: Maintaining the Directivity of Antenna Arrays, *Proc. IRE*, **22**, 847, July, 1934; **22**, 1313, November, 1934.
146. Kirby, Samuel S.: Radio Field Intensity and Distance Characteristics of a High Vertical Broadcast Antenna, *Proc. IRE*, **24**, 859, June, 1936.
147. Kleinwachter, H.: Determination of the Radiation Diagrams of Radiator Groups; Directive or Anti-fading Aerial Systems and Acoustic Radiators; an Electro-mechanical Method, abstract, *Wireless Engr.*, **19**, 576, December, 1942.
148. Labus, J. W.: Broadcast Antenna for Low Angle Radiation, *Proc. IRE*, **23**, 935, August, 1935.
149. Laport, E. A.: Directional Antenna Design, *Electronics*, **9**, 22, April, 1936.
150. Leeman, A.: Modified Protective Gap for Transmitting Antennas, *Electronics*, **16**, 128, May, 1943.
151. McKinnon, K. A.: Experimental Comparison of Shunt- and Series-excitation of a High Uniform Cross-section Vertical Radiator, *Can. J. Research*, November, 1939, p. 227.
152. McPherson, W. L.: Electrical Properties of Aerials for Medium and Long Wave Broadcasting, *Elec. Commun.*, April, 1938, p. 306.
153. Morrison, J. F., and P. H. Smith: Shunt-excited Antenna, *Proc. IRE*, **25**, 673, June, 1937.
154. ———: Simple Method for Observing Current Amplitude and Phase Relations in Antenna Arrays, *Proc. IRE*, **25**, 1310–1326, October, 1937.
155. ———: Simplifying Adjustments of Antenna Arrays, *Electronics*, January, 1940, p. 70.
156. Moulton, C. H.: Signal Distortion by Directional Broadcast Systems, *Proc. IRE*, **40**, 595, May, 1952.
157. Prestholdt, Ogden: WABC's New Two-way Antenna, *Electronic Inds.*, **4**, 88, February, 1945.
158. Roder, H.: Elimination of Phase Shifts between the Currents in Two Antennas, *Proc. IRE*, **22**, 374, March, 1934.
159. Rountree, J. G.: Calculator for Two-element Directive Arrays, *Proc. IRE*, **32**, 760, December, 1944.
160. Sinclair, D. B.: Impedance Measurements on Broadcast Antennas, *Communications*, **19**, 5, July, 1939.
161. Smith, Carl E.: Critical Study of Two Broadcast Antennas, *Proc. IRE*, **24**, 1329, October, 1936.
162. Smith, C. E., and E. L. Gove: Electromechanical Calculator for Directional Antenna Patterns, *Elec. Eng.*, **62**, *Trans.* 78, February, 1943.
163. Smith, C. E.: "Directional Antennas," Cleveland Institute of Radio Electronics, Cleveland, Ohio, 1946.
164. Spangenberg, K.: Charts for the Determination of the R.m.s. Value of Horizontal Radiation Pattern of Two-element Broadcast Arrays, *Proc. IRE*, **30**, 237, May, 1942.
165. Taylor, J. P.: Cathode-ray Antenna Phasemeter, *Electronics*, **12**, 62, April, 1939.
166. Wait, J. R., and W. A. Pope: The Characteristics of a Vertical Antenna with a Radial Conductor Ground System, *Convention Record IRE*, Part 1, p. 79, 1954.
167. Williams, H. P.: Machine for Calculating Polar Diagrams, *Electronics*, **16**, 196, September, 1943.

Broadcast, F-m, and TV Receiving Antennas:

168. Aceves, J. G.: Antenna System for FM Reception, *Electronics*, September, 1941.
169. Benham, W. E.: Aerial Coupling Systems for Television, *Wireless Engr.*, **15**, 555, October, 1938.
170. Carlson, W. L., and V. D. Landon: New Antenna Kit Design, *RCA Rev.*, **2**, 60, July, 1937.
171. ———: Television Reception with Built-in Antennas for Horizontally and Vertically Polarized Waves, *RCA Rev.*, **6**, 443, April, 1942.
172. Fruin, D.: Central Antenna System, *Electronics*, November, 1939, p. 37.
173. Goldman, Stanford: Shielded Loop for Noise Reduction in Broadcast Reception, *Electronics*, **2**, 20, October, 1938.
174. Landon, V. D., and J. D. Reid: A New Antenna Design for Noise Reduction, *Proc. IRE*, March, 1939, p. 188.
175. Morgan, H. K.: Multi-frequency Tuned Antenna System, *Electronics*, August, 1940, p. 42.
176. Mountjoy, G.: Measurement of Effective Height of Automobile Antennas, *RCA Rev.*, **3**, 369, January, 1939.
177. Noll, E. M.: Antennas for Television Receivers, *Radio News*, **33**, 40, May, 1945.
178. Rettenmeyer, F. X.: Radio-frequency Distributing Systems, *Proc. IRE*, **23**, 1286, November, 1935.
179. Seeley, S. W.: Effect of Receiving Antenna on Television Reception Fidelity, *RCA Rev.*, April, 1939, p. 433.
180. Sturley, K. A.: Receiver Aerial Coupling Circuits, *Wireless Engr.*, April, 1941, p. 137, and May, 1941, p. 190.
181. Taylor, S. G.: RCA Spiderweb Snags All Waves, *Radio News*, **18**, 292, November, 1936.
182. Wheeler, H. A., and V. E. Whitman: Design of Doublet Antenna Systems for All-wave Radio Reception, *Proc. IRE*, **24**, 1257, October, 1936.
183. New Form of Antenna; V-doublet System, *Electrician*, **114**, 824, June 21, 1935. (Same abr., *Gen. Elec. Rev.*, **38**, 395, August, 1935.)
184. Modern Antenna System for Apartment Houses, *Gen. Elec. Rev.*, May, 1939, p. 226.
185. Multicoupler Antenna System Includes FM Reception, *Rev. Sci. Instruments*, **13**, 86, February, 1942.

H-f Antennas:

186. Bailey, D. K., R. Bateman, and R. C. Kirby: Radio Transmission at VHF by Scattering and Other Processes in the Lower Ionosphere, *Proc. IRE*, **43**, 1181, October, 1955.
187. Barfield, R. H., and W. Ross: Short-wave Adcock Direction Finder, *J. IEE*, **81**, 683, 1937; *Wireless Sec. IEE*, **13**, 39, March, 1938.
188. Baumler, M., K. Kruger, H. Plendl, and W. Pfitzer: Radiation Measurements of a Short Wave Directive Antenna at the Nauen High Power Radio Station, *Proc. IRE*, May, 1931, p. 812.
189. Beverage, H. H., and H. O. Peterson: Diversity Receiving System of RCA Communications, Inc., for Radiotelegraphy, *Proc. IRE*, **19**, 545, April, 1931.
190. Brown, R.: Transoceanic Radiotelephone Development; Multiple Unit Steerable Antenna, *Proc. IRE*, **25**, 1131, September, 1937.
191. ———: Researches in Radiotelephony; Multiple-unit Steerable Antenna, *J. IEE*, **83**, 395, September, 1939.
192. Bruce, E.: Developments in Short-wave Directive Antennas, *Proc. IRE*, **19**, 1406, August, 1931; *Bell System Tech. J.*, **10**, 656, October, 1931.
193. ———, A. C. Beck, and L. R. Lowry: Horizontal Rhombic Antennas, *Proc. IRE*, **23**, 24, January, 1935.
194. ——— and ———: Experiments with Directivity Steering for Fading Reduction, *Proc. IRE*, **23**, 357, April, 1935; *Bell System Tech. J.*, **14**, 195, April, 1935.
195. Cafferata, H.: A Generalized Radiation Formula for Horizontal Rhombic Antennas, *Marconi Rev.*, Nos. 80, 81, and 82; **9** (1), January–March, 1946; **9** (2), April–June, 1946; **9** (3), July–September, 1946.
196. Carswell, I.: "Properties of Asymmetric Dipoles," Tech. Rept. 52, Stanford Research Institute, Menlo Park, Calif., December, 1955.
197. Carter, P. S., C. W. Hansell, and N. E. Lindenblad: Development of Directive Transmitting Antennas by RCA Communications, *Proc. IRE*, **19**, 1773, October, 1931.
198. Chamberlain, A. B.: CBS International Broadcast Facilities, *Proc. IRE*, **30**, 118, March, 1942.
199. Cleckner, D. C.: Effect of Feed on Pattern of Wire Antennas, *Electronics*, **20**, 103, August, 1947.
200. Everitt, W. L., and J. F. Byrne: Single Wire Transmission Lines for Short-wave Antennas, *Ohio State Univ., Eng. Exp. Sta. Bull.*, **52**, 1, 1930.
201. Foster, Donald: Radiation from Rhombic Antennas, *Proc. IRE*, **25**, 1327, October, 1937.
202. Foster, D. E., and C. W. Finnigan: Method of Measuring the Effectiveness of Electrostatic Loop Shielding, *Proc. IRE*, **31**, 253, June, 1943.
203. Fereday, R. A.: Sense Finding Device for Use with Spaced-aerial Direction Finders, *J. IEE*, **84**, 96, 1939.
204. Franz, K.: Gain and Absorption Area of Large Directional Aerials, *Hochfrequenztechn.*, December, 1939, p. 198.
205. Friis, H. T.: New Directional Receiving System, *Proc. IRE*, **13**, 685, December, 1925.
206. ——— and C. R. Feldman: Multiple Unit Steerable Antenna for Short-wave Reception, *Proc. IRE*, **25**, 841, July, 1937; *Bell System Tech. J.*, **16**, 337, July, 1937.
207. Gardiner, E. L.: Aerial Reflectors, *Wireless World*, October, 1940, p. 419.
208. Gill, A. J.: Developments in Radio Engineering Carried Out by the Post Office Engineering Department, *J. IEE*, February, 1939, p. 256.
209. Goddess, G.: NBC Beam Antennas, *Communications*, March, 1939, p. 16.
210. Goldman, S.: Dipoles and Reflectors, *Electronics*, May, 1940, p. 20.
211. Grosskopf, J., and K. Vogt: Polarization Measurements in the Field of Horizontal Transmitting Dipole, abstract, *Wireless Engr.*, **21**, 437, September, 1944.
212. Guertler, R.: Impedance Transformation in Folded Dipoles, *Proc. IRE*, **38**, 1042, September, 1950.
213. Hansen, W. W., and L. M. Hollingworth: Design of Flat-shooting Antenna Arrays, *Proc. IRE*, February, 1939, p. 137.
214. Harper, A. E.: "Rhombic Antenna Design," Van Nostrand, 1941.
215. Harrison, C. W., Jr.: Characteristics of the Two-antenna Array, *Proc. IRE*, **31**, 75, February, 1943.
216. ———: Radiation from Vee Antennas, *Proc. IRE*, **31**, 362, July, 1943.
217. ———: Radiation Field of Long Wires, with Application to Vee Antennas, *J. Applied Phys.*, **14**, 537, October, 1943.
218. Hayes, L. W., and B. N. McLarty: Empire Service Broadcasting Station at Daventry, *J. IEE*, September, 1939, p. 328.
219. Howe, G. W. O.: Polar Diagram of a Simple Broadside Array, *Wireless Engr.*, **19**, 193, May, 1942.
220. King, Ronold: Asymmetrically Driven Antennas and the Sleeve Dipole, *Proc. IRE*, **38**, 1154, October, 1950.
221. Kraus, J. D.: Multi-wire Dipole Antennas, *Electronics*, January, 1940, p. 26.
222. Lewin, L.: Rhombic Transmitting Aerial, *Wireless Engr.*, May, 1941, p. 140.
223. Linton, R. L., Jr.: The Folded Fan as a Broad-band Antenna, *Proc. IRE*, **39**, 1436, November, 1951.
224. Meissner, A.: Directional Radiation with Horizontal Antennas, *Proc. IRE*, **15**, 928, November, 1927.
225. ——— and H. Rothe: On the Determination of the Optimum Radiation Angle for Horizontal Antennas, *Proc. IRE*, **17**, 35, January, 1929.
226. Mushiake, Y.: An Exact Step-up Impedance-ratio Chart of a Folded Antenna, *Trans. IRE*, **AP-2**, 164, October, 1954.
227. Page, H.: Measured Performance of Horizontal Dipole Transmitting Arrays, *Electrician*, **133**, 408, Nov. 3, 1944; *Elec. Review (London)*, **135**, 625, Nov. 3, 1944.

228. Polkinghorn, F. A.: Single Sideband MUSA Receiving System for Commercial Operation on Transatlantic Radiotelephone Circuits, *Bell System Tech. J.*, April, 1940, p. 306.
229. Ross, W.: Calibration of 4-aerial Adcock Direction Finders, *J. IEE*, August, 1939, p. 192.
230. Siegel, E.: Reception with Short Antennas, *Electronics*, **9**, 48, January, 1936.
231. Smale, J. A.: Comparative Merits of Different Types of Directive Aerials for Communications, *J. IEE*, **91**, 3, 12, March, 1944; *Electrician*, **132**, 74, Jan. 28, 1944; *Elec. Review (London)*, **134**, 119, Jan. 28, 1944.
232. Sterba, E. J.: Theoretical and Practical Aspects of Directional Transmitting Systems, *Proc. IRE*, **19**, 1184, July, 1931.
233. Van Slooten, J.: Receiving Aerials, *Philips Tech. Rev.*, November, 1939, p. 320.
234. Walmsley, T.: Beam Arrays and Transmission Lines, *J. IEE*, February, 1931.
235. ———: New Type of Directive Aerial, *Wireless Engr.*, **9**, 662, November, 1932.
236. Wells, N.: Beam Wireless Telegraphy, *Elec. Review (London)*, **102**, 898, 940, May 24–June 1, 1928.
237. Wells, N.: Short-wave Dipole Aerials, *Wireless Engr.*, **20**, 219, May, 1943.
238. ———: Quadrant Aerial; An Omni-directional Wide-band Horizontal Aerial for Short Waves, *J. IEE*, **91**, Part 3, 182, December, 1944; Brief paper, *Marconi Rev.*, **9**, January–March, 1946.
239. Chart for Rhombic Antenna Design, *Electronic Inds.*, **3**, 112, October, 1944.
240. Marconi Series-phase Aerial, *Engineering*, **138**, 235, Aug. 31, 1934.
241. Scatter Propagation Issue, *Proc. IRE*, **43**, October, 1955.

V-h-f and U-h-f Antennas:

242. Alford, A.: Long Slot Antennas, *Proc. Natl. Elect. Conf.*, Chicago, Ill., 1946.
243. Alford, A., and A. G. Kandoian: UHF Loop Antennas, *Elec. Eng.*, **59**, 843–848, April, 1940; *Elec. Commun.*, **18**, 255–265, April, 1940.
244. ———: Ultrashort Electromagnetic Waves; Design of Radiators or Antennas for Ultrashort Waves, *Elec. Eng.*, **62**, 303, 338, July–August, 1943.
245. Bailin, L. L.: The Radiation Field Produced by a Slot in a Large Circular Cylinder, *Trans. IRE*, **AP-3**, 128, July, 1955.
246. Balin, L. L., R. S. Wehner, and I. P. Kaminow: Empirical Approximations to the Current Values for Large Dolph-Tchebyscheff Arrays, *Convention Record IRE*, Part I, Antennas and Propagation, p. 3, 1954.
247. Barbiere, D.: A Method for Calculating the Current Distribution of Tchebyscheff Arrays, *Proc. IRE*, **40**, 78, January, 1952 (see also *Proc. IRE*, **40**, 991, August, 1952; **42**, 1021, June, 1954).
248. Barfield, R. H., and R. A. Fereday: Improved Medium Wave Adcock Direction Finder, *Wireless Sec., IEE*, **13**, 33, March, 1938.
249. Barrow, W. L., and F. M. Greene: Rectangular Hollow Pipe Radiators, *Proc. IRE*, December, 1938, p. 1498.
250. Barrow, W. L., and L. J. Chu: Theory of the Electromagnetic Horn, *Proc. IRE*, January, 1939, p. 51.
251. ——— and F. D. Lewis: Sectoral Electromagnetic Horn, *Proc. IRE*, January, 1939, p. 41.
252. ———, L. J. Chu, and Jansen: Biconical Electromagnetic Horns, *Proc. IRE*, December, 1939, p. 769.
253. ——— and C. Shulman: Multi-unit Electromagnetic Horns, *Proc. IRE*, March, 1940, p. 130.
254. Bonnert, J. I.: Measurements on Elliptically Polarized Antennas, *Proc. IRE*, **39**, 549, May, 1951.
255. Booker, H. G.: Slot Aerials and Their Relation to Complementary Wire Aerials (Babinet's Principle), *J. IEE*, **93**, Part III-A, 620, 1946.
256. Bromel, R.: Beam Properties of Small Parabolic Reflectors with Various Excitations, *Hochfrequenztechn.*, **48**, 81, 121, 1936.
257. Brown, G. H.: Turnstile Antenna for UHF Broadcasting, *Electronics*, April, 1936.
258. ——— and J. Epstein: UHF Antenna of Simple Design, *Communications*, July, 1940, p. 3.
259. ——— and ———: Pretuned Turnstile Antenna, *Electronics*, **18**, 102, June, 1945.
260. Brown, G. H., and Woodward: Circularly-polarized Omnidirectional Antenna, *RCA Rev.*, **8**, 259, June, 1947.
261. Builder, G.: Duplex Operation on a Single Aerial, *Electronics*, February, 1940, p. 61.
262. Bystrom, A., Jr., and D. G. Berntsen: An Experimental Investigation of Cavity-mounted Helical Antennas, *Trans. IRE*, **AP-4**, 53, January, 1956.
263. Caporale, Peter: CAA Instrument Landing System, *Electronics*, March, 1945 (excellent bibliography on blind-landing systems); February, 1945, p. 116, and March, 1945, p. 128.
264. ———: CAA-RTCA Instrument Landing Installation, *Electronic Inds.*, March, 1944, p. 92.
265. Carter, P. S.: Simple Television Antennas, *RCA Rev.*, October, 1939, p. 168.
266. ———: Antenna Arrays around Cylinders, *Proc. IRE*, **31**, 671, December, 1943.
267. Chu, L. J., and W. L. Barrow: E-m Waves in Hollow Metal Tubes of Rectangular Cross-section, *Proc. IRE*, December, 1938, p. 1520.
268. ——— and ———: Electromagnetic Horn Design, *Elec. Eng.*, July, 1939, p. 333.
269. ———: Calculation of Radiation Properties of Hollow Pipes and Horns, *J. Applied Phys.*, September, 1940, p. 603.
270. Deschamps, G. A.: Geometrical Representation of the Polarization of a Plane Electromagnetic Wave, *Proc. IRE*, **39**, 540, May, 1951.
271. Dolph, C. L.: A Current Distribution for Broadside Arrays Which Optimizes the Relationship between Beamwidth and Side-lobe Level, *Proc. IRE*, **34**, 335, June, 1946.
272. Foster, Donald: Loop Antennas with Uniform Current, *Proc. IRE*, October, 1944, pp. 603–607.

273. Franklin, J. C.: Installation of Antennas for Direction Finding, *Air Commerce Bull.*, **9**, 256, April, 1938.
274. Hansell, C. W.: Radio Relay Systems Developed by the Radio Corporation of America; Couplings between Directional Antennas, *Proc. IRE*, **33**, 164, March, 1945.
275. Hasenbeck, H. W.: Design Data for Ground Plane Antennas, *Electronics*, **16**, 98, August, 1943.
276. Hatkin, L.: Elliptically Polarized Waves, *Proc. IRE*, **38**, 1455, December, 1950.
277. Haycock, O. C., and F. L. Wiley: Radiation Patterns and Conductance of Slotted-cylinder Antennas, *Proc. IRE*, **40**, 349, March, 1952.
278. ——— and J. S. Ajioka: Radiation Characteristics of Helical Antennas of Few Turns, *Proc. IRE*, **40**, 989, August, 1952.
279. Herold, E. W., and L. Malter: Some Aspects of Radio Reception at Ultra-high Frequency; The Antenna and the Receiver Input Circuits, *Proc. IRE*, **31**, 423, August, 1943.
280. Holtz, R. F.: The Pylon Antenna, *Broadcast News*, **44**, 36, October, 1946.
281. Hromada, J. C.: Ultra-high Frequency Antennas in Aviation Radio, *Radio News*, **27**, 14, May, 1942.
282. Jordan, E. C., and W. E. Miller: Slotted-cylinder Antennas, *Electronics*, **20**, 90, February, 1947.
283. Kales, M. L.: Elliptically Polarized Waves and Antennas, *Proc. IRE*, **39**, 544, May, 1951.
284. Kandoian, A. G.: Radiating Systems and Wave Propagation, *Electronics*, **15**, 39, April, 1942.
285. ———: 3 New Antenna Types and Their Applications, *Elec. Commun.*, March, 1946, pp. 27–34; also *Proc. IRE*, **34**, 70W–75W, February, 1946.
286. Kenyon, R. W.: Principles of Loran in Position Location, *Electronic Inds.*, December, 1945, p. 106.
287. ———: Loran Transmitting Stations, *Electronics*, March, 1946, p. 109.
288. Kraus, J. D.: Antenna Arrays with Close-spaced Elements, *Proc. IRE*, February, 1940, p. 76.
289. ———: Corner-reflector Antenna, *Proc. IRE*, November, 1940, p. 513.
290. Lange, P.: Measurements on Dipoles in the Decimetric Region, *Wireless Engr.*, November, 1941, p. 465.
291. Laport, Edmund A., and James B. Knox: Radiating System for 75 Megacycle Cone-of-silence Marker, *Proc. IRE*, **30**, 26, January, 1942.
292. Lesensky, M. M.: UHF Aerial Coupling Circuits, *Indian J. Phys.*, December, 1940, p. 67.
293. Lewis, R. F.: Vertical Turnstile Antenna, U.S. Patent 2210491.
294. Lindenblad, N. E.: Television Transmitting Antenna for Empire State Bldg., *RCA Rev.*, April, 1939, p. 387.
295. Masters, R. W.: A Power-equalizing Network for Antennas, *Proc. IRE*, **37**, 735, July, 1949.
296. McPherson, W. L., and E. H. Ulrich: Micro-ray Communication, *J. IEE*, **78**, 632, June, 1936.
297. Morgan, H. K.: Beacon Antenna Characteristics, *Air Commerce Bull.*, **9**, 77, October, 1937.
298. Moullin, E. B.: Theory and Performance of Corner Reflectors for Aerials; abstracts, *Electrician*, **133**, 408, Nov. 3, 1944; *Elec. Review (London)*, **135**, 625, Nov. 3, 1944.
299. Nagy, A. W.: Experimental Study of Parasitic Wire Reflectors on 2.5 Meters, *Proc. IRE*, **24**, 233, February, 1936.
300. Naschke, E. S.: Police Satellite System; FM Signals Automatically Relayed to Central Point by Two Mountain-top Stations; Design of 60-degree Corner Reflector Receiving Antennas, *Electronics*, **17**, 94, May, 1944.
301. Palmer, L. S., and K. G. Gillard: Distribution of Ultra-high-frequency Currents in Long Transmitting and Receiving Antennae, *J. IEE*, **83**, 415, September, 1938.
302. Parker, R. H.: Coaxial Antenna for 112 Mc., *QST*, **29**, 40, June, 1945.
303. Polydroff, W. J.: Iron-core Loop Antennas, *Electronic Inds.*, **2**, 84, November, 1943.
304. Reber, G.: Electromagnetic Horns, *Communications*, February, 1939, p. 13.
305. Riblet, H. J.: Microwave Omnidirectional Antennas, *Proc. IRE*, **35**, 474, May, 1947.
306. ——— and C. L. Dolph: Discussion on "A Current Distribution for Broadside Arrays Which Optimizes the Relationship between Beam Width and Side-lobe Level," *Proc. IRE*, **35**, 489, May, 1947.
307. Rumsey, V. H.: Transmission between Elliptically Polarized Antennas, *Proc. IRE*, **39**, 535, May, 1951 (see also *Proc. IRE*, **43**, **733**, June, 1955).
308. Schafer, O.: Comparison between Electric Horns and Other Directional Radiators, *Wireless Engr.*, **22**, 90, February, 1945.
309. Scheldorf, M. W.: FM Circular Antenna, *Gen. Elec. Rev.*, **46**, 163, March, 1943.
310. ———: New Studio-to-transmitter Antenna, *Proc. IRE*, **33**, 106, February, 1945.
311. Seeley, S. W.: Shoran Precision Radar, *Elec. Eng.*, April, 1946, p. 232.
312. Sherman, J. B.: Circular Loop Antennas at Ultra-high Frequencies, *Proc. IRE*, **32**, 534, September, 1944.
313. Sinclair, G.: The Patterns of Slotted-cylinder Antennas, *Proc. IRE*, **36**, 1487, December, 1948.
314. ———: Patterns of Antennas Located near Cylinders of Elliptical Cross Section, *Proc. IRE*, **39**, 660, June, 1951.
315. ———: The Transmission and Reception of Elliptically Polarized Waves, *Proc. IRE*, **38**, 148, February, 1950.
316. Smith-Rose, R. L., and H. G. Hopkins: Radio Direction Finding on Wavelengths between 6 and 10 Meters, *J. IEE*, **83**, 87, 1938; *Wireless Sec., IEE*, **13**, 303, September, 1938.
317. ——— and H. G. Hopkins: Radio Direction-finding on Wavelengths between 2 and 3 Meters, *J. IEE*, August, 1940, p. 154.
318. Southworth, G. C., and A. P. King: Metal Horns as Directive Receivers of Ultra-short Waves, *Proc. IRE*, February, 1939, p. 95.
319. Stahl, C. I. H. A.: Full Parabolic Reflectors for Microwaves, *Philips Transmitting News*, **3**, 14, 1937.

320. **Stegen, R. J.: Excitation Coefficients and Beamwidths of Tschebyscheff Arrays,** *Proc. IRE*, **41,** 1671, November, 1953.
321. Taylor, J. P.: Square-loop F-m Antenna at WBRL, *Electronics*, **18,** 96, March, 1945.
322. Uda, S.: High Angle Radiation of Short Electric Waves, *Proc. IRE*, **15,** 377, May, 1927.
323. Watts, C. B.: Simultaneous Radiation of Odd and Even Patterns by a Linear Array, *Proc. IRE*, **40,** 1236, October, 1952.
324. Wheeler, H. A.: A Helical Antenna for Circular Polarization, *Proc. IRE*, **35,** 1484, December, 1947.
325. Williams, E. M.: Radiation Characteristics of Short-wave Loop Aerials, *Proc. IRE*, October, 1940, p. 480.
326. Wolff, Irving: Determination of the Radiating System Which Will Produce a Specified Directional Characteristic, *Proc. IRE*, **25,** 630, May, 1937.
327. Wong, J. Y.: Radiation Conductance of Axial and Transverse Slots in Cylinders of Elliptical Cross Section, *Proc. IRE*, **41,** 1172, September, 1953.
328. ———: Radiation Patterns of Slotted-elliptic Cylinder Antennas, *Trans. IRE*, **AP-3,** 200, October, 1955.

S-h-f Antennas:

329. Beam, R. E., M. M. Astrahan, and H. F. Mathis: Open-ended Waveguide Radiators, *Proc. Natl. Elect. Conf.*, **4,** 472, 1948.
330. Braun, E. H.: Gain of EM Horns, *Proc. IRE*, **41,** 109, January, 1953.
331. ———: Some Data for the Design of Electromagnetic Horns, *Trans. IRE*, **AP-4,** 29, January, 1956.
332. Butterfield, F. E.: Dielectric Sheet Radiators, *Trans. IRE*, **AP-2,** 152, October, 1954.
333. Carlson, J. F., and A. E. Heins: The Reflection of an Electromagnetic Plane Wave by an Infinite Set of Plates, I, *Quart. Applied Math.*, **4,** 313, January, 1947.
334. Carter, D.: Diffuse Radiation in Pencil Beam Antennas, *Convention Record IRE*, Part I, Ant. and Prop., p. 60, 1954.
335. Chlavin, A.: A New Antenna Feed Having Equal E- and H-plane Patterns, *Trans. IRE*, **AP-2,** 113, July, 1954.
336. Cohn, S. B.: Analysis of Metal-strip Delay Structure for Microwave Lenses, *J. Applied Phys.*, **20,** 257, March, 1949.
337. ———: Microwave Measurements on Metallic Delay Media, *Proc. IRE*, **41,** 1177, September, 1953.
338. ———: Artificial Dielectrics for Microwaves, *Proc. Symp. Modern Advances in Microwave Techniques*, Polytechnic Institute of Brooklyn, 1954.
339. Corkum, R. W.: Isotropic Artificial Dielectric, *Proc. IRE*, **40,** 574, May, 1952.
340. Cullen, A. L.: On the Channel Section Waveguide Radiator, *Phil. Mag.*, **40,** 417, April, 1949.
341. Cutler, C. C.: Parabolic-antenna Design for Microwaves, *Proc. IRE*, **35,** 1284, November, 1947.
342. DeVore, H. B., and H. Iams: Microwave Optics between Parallel Conducting Plates, *RCA Rev.*, **9,** 721, December, 1948.
343. Ehrlick, M., and L. Newkirk: Corrugated Surface Antennas, *Convention Record IRE*, Part 2, Ant. and Communications, p. 18, 1953.
344. ——— and Joann Short: Mutual Coupling Considerations in Linear-slot Array Design, *Proc. IRE*, **42,** 956, June, 1954.
345. Elliott, R. S.: On the Theory of Corrugated Plane Surfaces, *Trans. IRE*, **AP-2,** 71, April, 1954.
346. Goatley, C., and C. F. Parker: Symmetrical Microwave Lenses, *Convention Record IRE*, Part 1, Ant. and Prop., p. 13, 1955.
347. Hannan, P. W.: Reflections in Microwave Antennas and Their Harmful Effects, *Convention Record IRE*, Part 1, Ant. and Prop., p. 39, 1954.
348. Hansen, W. W.: "Radiating Electromagnetic Waveguides," U.S. Patent 2402622.
349. Hines, J. N., V. H. Rumsey, and C. H. Walter: Traveling-wave Slot Antennas, *Proc. IRE*, **41,** 1624, November, 1953.
350. Holzschuk, D. L.: The NRL Precision "Big-Dish" Antenna, *Convention Record IRE*, Part I, Ant. and Prop., p. 32, 1955.
351. Jakes, W. C.: Gain of Electromagnetic Horns, *Proc. IRE*, **39,** 160, February, 1951.
352. Jones, E. M. T., and R. A. Folsom, Jr.: A Note on the Circular Dielectric-disk Antenna, *Proc. IRE*, **41,** 798, June, 1953.
353. ———: Paraboloid Reflector and Hyperboloid Lens Antennas, *Trans. IRE*, **AP-2,** 119, July, 1954.
354. Kelly, K. C., and R. S. Elliot: Serrated Waveguide Theory and Experiment, *Convention Record IRE*, Part I, Ant. and Prop., p. 6, 1955.
355. King, A. P.: The Radiation Characteristics of Conical Horn Antennas, *Proc. IRE*, **38,** 249, March, 1950.
356. Kock, W. E.: Metal-lens Antennas, *Proc. IRE*, **34,** 828, November, 1946.
357. ———: Metal Delay Lenses, *Bell System Tech. J.*, **27,** 58, January, 1948.
358. ———: Path-length Microwave Lenses, *Proc. IRE*, **37,** 852, August, 1949.
359. Korman, N. I., and J. R. Ford: An Achromatic Microwave Antenna, *Proc. IRE*, **38,** 1455, December, 1950.
360. Lawson, J. D.: The Practical Aspects of Paraboloid Aerial Design, *J. IEE*, **93,** Part IIIA, 1511, 1946.
361. Luneberg, R. K.: "Mathematical Theory of Optics," Brown University Press, 1944.
362. Morita, T., and S. B. Cohn: Microwave Lens Matching by Simulated Quarter-wave Transformers, *Trans. IRE*, **AP-4,** 33, January, 1956.
363. Mueller, G. E., and W. A. Tyrrell: Polyrod Antennas, *Bell System Tech. J.*, **26,** 837, October, 1947.

364. Peeler, G. E. M., and D. H. Archer: A Two-dimensional Microwave Luneberg Lens, *Trans. IRE*, **AP-1,** 12, July, 1953.
365. ———, K. S. Kelleher, and H. P. Coleman: Virtual Source Luneberg Lenses, *Trans. IRE*, **AP-2,** 94, July, 1954.
366. Pippard, A. B.: The Hoghorn—an Electromagnetic Horn Radiator of Medium-sized Aperture, *J. IEE*, **93,** Part IIIA, 1536, March–May, 1946.
367. ——— and N. Elson: The Elimination of Standing Waves in Aerials Employing Parabolic Reflectors, *J. IEE*, **93,** Part IIIA, 1531, March–May, 1946.
368. Puro, W. O., and K. S. Kelleher: Isotropic Variable Index Media, *Convention Record IRE*, Part I, Ant. and Prop., p. 76, 1954.
369. Rhodes, D. R.: An Experimental Investigation of the Radiation Patterns of Electromagnetic Horn Antennas, *Proc. IRE*, **36,** 1101, September, 1948.
370. Rotman, W.: "The Channel Guide Radiator," Air Force Cambridge Research Center, Report E5054, 1950.
371. ———: A Study of Single-surface Corrugated Guides, *Proc. IRE*, **39,** 952, August, 1951.
372. Rumsey, V. H.: The Design of Frequency Compensating Matching Sections, *Proc. IRE*, **38,** 1191, October, 1950.
373. Ruze, J.: Wide-angle Metal-plate Optics, *Proc. IRE*, **38,** 53, January, 1950.
374. Stengen, R. J., and R. H. Reed: Arrays of Closely-spaced Nonresonant Slots, *Trans. IRE*, **AP-2,** 109, July, 1954.
375. Stephenson, B. T., and C. H. Walter: Endfire Slot Antennas, *Trans. IRE*, **AP-3,** 81, April, 1955.
376. Stuetzer, O. M.: Development of Artificial Microwave Optics in Germany, *Proc. IRE*, **38,** 1053, September, 1950.

CHAPTER 21

BROADCASTING

By Carl G. Dietsch[1]

1. Principal Elements of Broadcasting System. All the equipment of a broadcasting system extending from the microphone to the radiating antenna of the radio transmitting station will be considered as part of the system. A general circuit layout of typical facilities of the kind used in the larger broadcasting centers for supplying a network of stations with program service is represented by the simplified diagram, Fig. 1. Equipment of a single studio is represented; that of other studios of the usual group would be similar and would be at the point marked on the program bus. Inasmuch as many programs, such as the broadcasting of special events, originate at remote points, in most cases a great distance from the studio, the layout of the facilities for remote pickups, sometimes termed "nemo" programs, has been included to illustrate the use of telephone lines as well as point-to-point radiotelephone communication to complete the circuits necessary.

A list of the essential elements of the system is as follows:

1. Microphones
   - *a.* Studio
   - *b.* Remote pickups
2. Apparatus for controlling and conveying microphone output
   - *a.* Studio control booth
     - (1) Preamplifiers
     - (2) Microphone mixers
     - (3) Studio amplifier
     - (4) Volume control or faders
     - (5) Volume indicator
     - (6) Monitoring speaker
   - *b.* Remote pickups
     - (1) Preamplifiers
     - (2) Volume controls or faders
     - (3) Volume indicator
     - (4) Monitoring equipment
     - (5) Radiotelephone or wire-line facilities for intercommunication
3. Master control-room apparatus
   - *a.* Volume controls
   - *b.* Studio amplifiers
   - *c.* Relays and switching apparatus
   - *d.* Network channel amplifiers
   - *e.* Volume indicator
   - *f.* Monitoring facilities
4. Telephone-line facilities to local radio transmitting stations and to distant radio transmitters connected to networks
5. Radio transmitter
   - *a.* Limiting amplifier
   - *b.* Volume controls
   - *c.* Volume indicator
   - *d.* Radio transmitter
   - *e.* Monitoring equipment
     - (1) Monitoring rectifier and speaker
     - (2) Modulation-percentage indicator
     - (3) Carrier-frequency monitor
   - *f.* Antenna

[1] Engineering Department, RCA Communications, Inc.

2. A-f Range. Perfect reproduction of a sound transmitted through an electro-acoustic system requires that the system pass all the audible frequencies of the sound in their relative intensities. Under these conditions of reproduction, the listener would be conveyed acoustically from his loud-speaker to a point near the sound source

FIG. 1. Schematic diagram of typical broadcast transmitting system.

except that the absolute sound level might be different from that experienced at the point of origin.

A correlated acoustic chart of the frequency range of various musical instruments within the orchestral range and the different voices which constitute the vocal range is shown in Fig. 2. The shaded keys are not included on a standard piano keyboard. The extreme organ range not shown on the chart is from 16 to 16,384 cps physical pitch. The extreme frequency-transmission ranges necessary to produce perfect

FIG. 2. Correlated acoustic chart showing the scientific or philosophical scale generally used by physicists and the international equally tempered scale based on $A = 435$ complete vibrations per second. The current orchestra or symphony scale based on $A = 440$ is generally used by musicians.

naturalness of speech and orchestral music are shown in Fig. 3. These ranges extend considerably above those of Fig. 2 because they include overtones and noise accompaniment additional to the fundamental tones. Research[1] indicates that for perfect reproduction of speech and music a frequency range between 30 and 15,000 cps is

[1] Snow, W. B., *J. Acoust. Soc. Am.*, July, 1931, p. 61.

desirable in order that the average ear may appreciate fully all the frequencies produced by the sound sources.

The curves shown in Fig. 4 are an indication of the relative qualities of reproduced orchestral music the frequency range of which was limited by electrical filters. It is apparent that, where a transmission system has a limited frequency range, such as

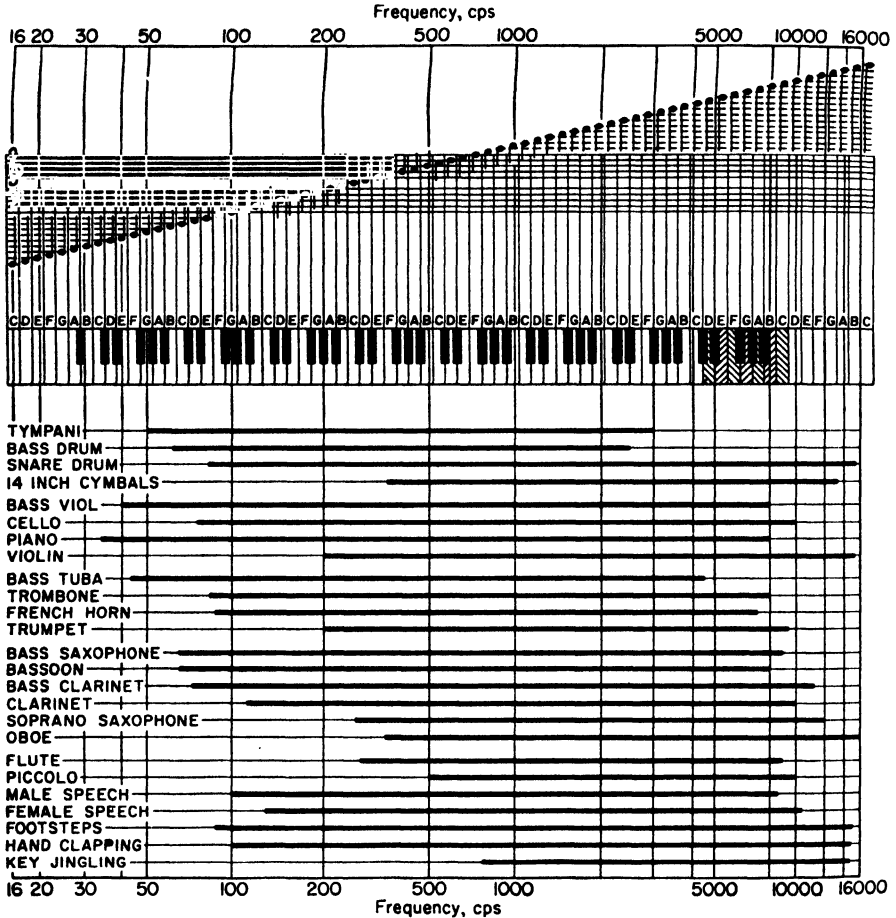



FIG. 3. Frequency range required for the reproduction of musical instruments, voice, and noise without noticeable distortion.

that which exists in broadcasting techniques, acceptable reproduction may be secured within a band width of between 30 and 9,000 cps.

The engineering and economic limitations of the frequency range used for standard broadcasting lie in restrictions on the use of the upper audio frequencies due largely to a limited band width of the modulation spectrum contained between the presently assigned carrier frequencies of 10-kc separation[1] for standard broadcasting in the spectrum between 550 and 1,600 kc.

[1] Eckersley, P. P., Minimum Frequency Separation, *Proc. IRE*, February, 1933, p. 195.

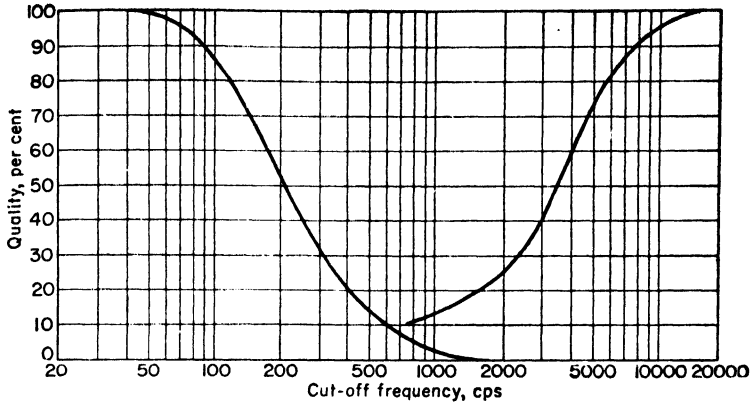



FIG. 4. Quality of orchestra music as a function of cutoff frequency.

Note	Cps	Organ pipe	Remarks
C_{10}	33,488	....	Beyond limit of audibility for average person
C_9	16,744		
	15,000	....	Considered ideal upper limit for perfect transmission of speech and music
	10,000	....	Considered as upper limit for high-quality transmission of speech and music
	9,000	....	Considered as satisfactory upper limit for high-quality transmission of speech and music
C_8	8,372	¾ in.	Highest note on fifteenth stop
C_7	4,186	....	Highest note of pianoforte
G_6	3,136		
E_8	2,673.2	....	Approximate resonant frequency of ear cavity
	3,000	....	Considered as satisfactory upper limit for transmission of speech for ordinary communication
C_6	2,093		
	2,000	....	Maximum sensitivity of human ear
	1,500	....	Mean speech frequency from articulation standpoint
	1,000	....	Representative frequency of telephone currents
E_4	659.3		
A_3	440	....	Orchestral tuning (see note below)
C_3	261.6		
C_2	130.8		
	100	....	Considered as satisfactory lower limit of good transmission of speech and music
E_2	82.4		
C_1	65.4	8 ft	Lowest note of cello
B_0	61.7		
C_0	32.7	16 ft	Lowest note of average church organ
	30	....	Considered ideal lower limit for perfect transmission of speech and music
A	27.5	....	Lowest note of pianoforte
G	24.5		
C	16.35	....	Lowest audible sound. Longest pipe of largest organ

Notes of the "gamut"	C	D	E	F	G	A	B	C
Vibration frequencies proportional to	1	$\frac{9}{8}$	$\frac{5}{4}$	$\frac{4}{3}$	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{15}{8}$	2
Intervals between successive notes	$\frac{9}{8}$	$\frac{10}{9}$	$\frac{16}{15}$	$\frac{9}{8}$	$\frac{10}{9}$	$\frac{9}{8}$	$\frac{16}{15}$	

FIG. 5. Frequencies to be transmitted on a high-quality system. NOTE: Nearest note is indicated. Scale A = 440 cps based on middle C_1 (symphony pitch) = 261.6 cps.

An overlapping of the modulation frequencies of a "wanted" station by those of an "unwanted" station of 10,000-cps separation restricts the range of frequencies to a broadcast listener usually considerably below that passed by the broadcasting system itself. The high quality of programs available from broadcasting facilities which have an over-all uniform frequency response from the microphone to the antenna within 2 db from 30 to 9,000 cps and above cannot be appreciated by the average listener because of limitations in the average receiver frequency response and restrictions in the present standard broadcast band due to the 10-kc channel separation. Adjacent channel interference in the form of cross talk or "monkey chatter" prevents the satisfactory reception of the higher audio frequencies.

This limitation is not apparent in the case of f-m broadcasting, since the channels are assigned sufficiently far apart to permit transmission and reception of audio frequencies up to 15,000 cps. This is obviously an advantage over standard broadcasting in that the entire audible frequency range may be satisfactorily transmitted and reproduced at the receiver.

Table 1. Peak Power of Musical Instruments

Instrument	Peak Power, Watts
Heavy orchestra	70
Large bass drum	25
Pipe organ	13
Snare drum	12
Cymbals	10
Trombone	6
Piano	0.4
Trumpet	0.3
Bass saxophone	0.3
Bass tuba	0.2
Bass viol	0.16
Piccolo	0.08
Flute	0.06
Clarinet	0.05
French horn	0.05
Triangle	0.05

3. Volume Range. Table 1 gives the peak power of various musical instruments playing triple forte. A violin playing very softly has an output of about 4 mw, whereas that of a full orchestra has a peak value of 70 watts. The intensity range of the sound sources in this case is about 43 db. Owing to limitations in broadcasting circuits, background noise, and the modulation capabilities of the transmitter, this volume range must be in most cases compressed within the limits which can be handled by the wire lines and their associated equipment, as well as the transmitter where serious amplitude distortion results if modulation peaks, except those of extremely short time duration, exceed the modulation capabilities of the transmitter.

STANDARD REFERENCE LEVELS

4. Vu and Decibels. The electrical signal level of an audio signal passing through each particular circuit of the broadcasting system, including the studio equipment, wire-line facilities, and the broadcasting station, must at all times be carefully adjusted. The adjustment must be such that the transmitter program signals will remain within the limits which every part of the system can transmit without objectionable distortion due to overloading or from interference caused by noise, such as the interference produced by cross talk, induction, rectifier ripple, etc., inherent in equipment as well as associated wire lines. A convenient and consistently accurate method of measuring the amplitude of the signals is required, as well as a reference level common to the entire system. On this reference level, termed the *zero reference point* or zero vu (volume units), is based the amplitude of the program waves throughout the system. The zero reference level of 1 mw is expressed as zero *dbm* for a single sine-wave frequency. It also serves as a reference level from which the amplitude of interference or noise may be measured in terms of a sine wave.

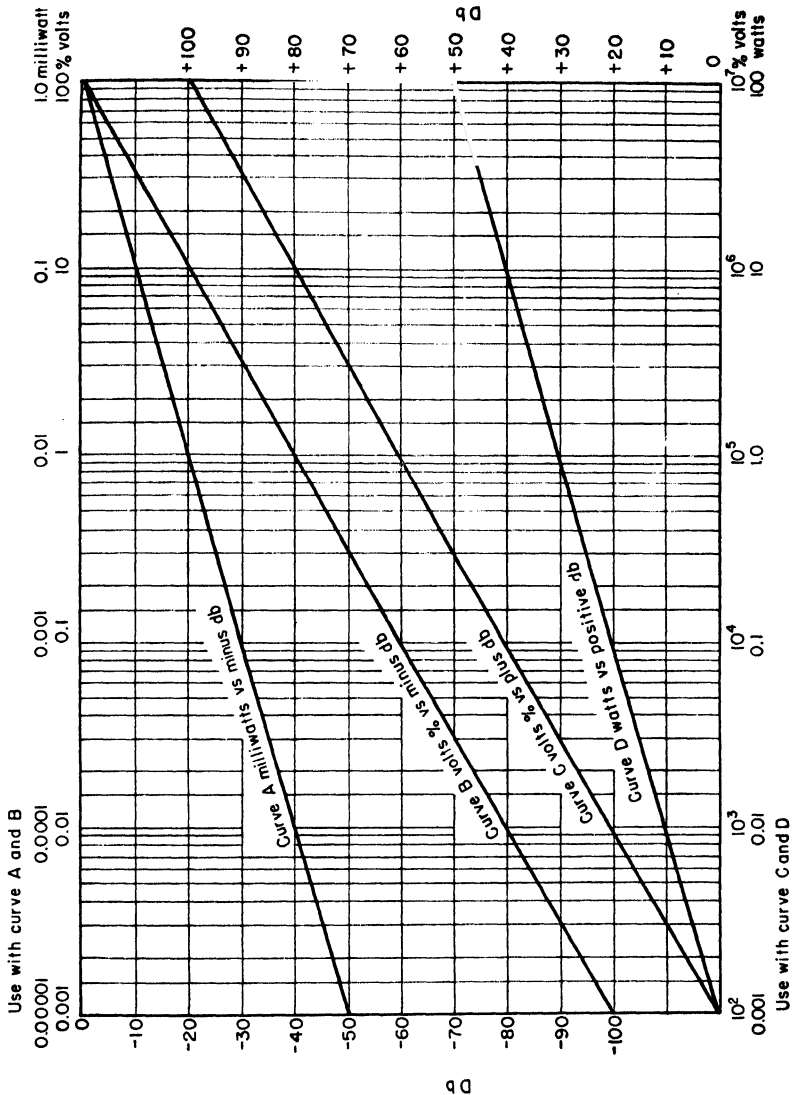



FIG. 6. Decibels, watts, and per cent volts; zero level is 1 mw across 600 ohms.

For broadcasting technique together with the interconnecting wire lines between studios and broadcasting stations there has been established a standard energy reference level of 1 mw. For the standard line impedance or pure resistance of 600 ohms at the terminals of a piece of apparatus in the system the zero reference level in vu would correspond to $\sqrt{0.6}$ rms volt of 1,000-cps sine-wave electrical energy as measured by a standard a-c voltmeter across the terminals (see also Chap. 13).

Since program signals have very complex wave shapes and because peaks of these complex waves are liable to cause overloading, a standard volume indicator[1] has

[1] Chinn, H. A., D. K. Gannett, and R. M. Morris, A New Standard Volume Indicator and Reference Level, *Proc. IRE*, January, 1940.

been adopted for the purpose of measuring program levels at all parts of a broadcasting system so that the correct signal level can be maintained without objectionable overloading. This instrument (see Sec. 21) is calibrated to read vu on a logarithmic scale. It has electrical characteristics approximately equivalent to those of an rms instrument. For signals having sinusoidal wave shape, the vu readings on this instrument should follow the decibel-voltage curve shown in Fig. 6. However, since the instrument is designed and used for measurement of complex program waves, the vu level of a particular program wave is as indicated by this standard volume indicator because of its particular characteristics. The term vu is therefore associated with the reading of this meter whereas the term decibel follows steady-state conditions and mathematical laws.

FIG. 7. Directional characteristics of 44BX velocity microphone.

Simultaneously with the establishment of the reference level of 1 mw, the former 0-db levels of 6, 10, 12½ mw, etc., were abandoned in 1939 and the value for standard apparatus and telephone-line termination impedances for broadcasting was changed to the present standard of 600 ohms.

AUDIO FACILITIES

5. Microphone Requirements. By means of the microphone, acoustic energy of sound waves is converted into electric energy, the wave shape of the latter conforming to that of the sound waves. The principal requirements of a microphone which will produce high-quality conversion are as follows: a relatively high sensitivity with respect to its inherent noise level, a uniform wave response over the frequency range desired, a substantially uniform frequency response over the angles included by its directivity characteristic, and mechanical and electrical ruggedness.

With some reservation, one may say that all forms of acoustoelectric transducers require the introduction of an obstacle into the path of the sound waves. To be effective, either the active element of a microphone must partake of or otherwise influence the motion of the air particles, or it must respond in some way to the pressure variations on its surface. Some portion of the instrument such as the outside case, regarded as a rigid obstacle, must reflect some of the incident wave energy, whereas the element which responds to vibration from the sound waves must reradiate some of the energy exciting it. An instrument of high sensitivity and efficiency must, therefore, absorb a considerable proportion of the sound energy reaching it and convert it into electric energy. Faithful reproduction, therefore, is dependent upon the physical size and shape of a microphone. These features enter into the distortion of the true sound field, as well as the characteristics of the elements used to convert the sound into electric energy with a minimum of wave distortion.

Inasmuch as the quality of reproduction of speech and music is dependent upon the acoustic properties of the room[1] containing the sound sources and the placement of the microphone with respect to them, satisfactory results using even the best instruments require a knowledge of the technique of microphone placement.

[1] Hanson, O. B., and R. M. Morris, Design and Construction of Broadcast Studios, *Proc. IRE*, **19**, January, 1931.

6. Velocity Microphone. This instrument gets its name from the movement of a metallic ribbon under the motion of air particles impinging upon it, thus setting up by electromagnetic induction an emf corresponding to the amplitude variations of an incident sound wave.

The commercial form of the RCA type 44BX[1] consists of a thin metallic ribbon suspended between the poles of powerful permanent magnets with the ribbon length perpendicular to, and its width in the plane of, the magnetic lines of force. It is moved from its position of equilibrium by the difference of pressure between its two sides. This pressure difference between the front and back of the ribbon is the same as that produced in a sound field between two points in space separated by this distance. The pressure difference between the front and back of the ribbon is proportional to frequency. Since the acoustic impedance of the system is also proportional to frequency

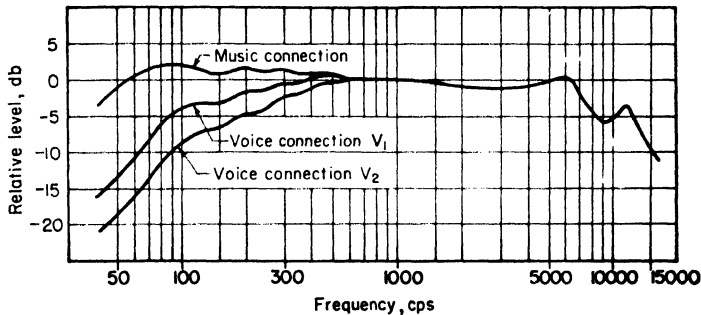



FIG. 8. Open-circuit frequency response of 44BX microphone.

and since velocity in a mechanical system is the ratio of the pressure to the acoustic impedance, the velocity of the ribbon is independent of frequency.

With a ribbon constructed to have a natural period below the audible range, the frequency response is free from severe irregularities prominent in some pressure-operated types because of cavity and diaphragm resonance and from pressure-doubling effects produced at the higher frequencies. The ribbon is made light enough so that its motion will conform with the motion of air particles even at very high frequencies, with a result that the response of the velocity microphone is uniform over a wide range of frequencies.

The velocity-type microphone is markedly directional. With a plane-progressive wave the response in front and back of the instrument varies with the cosine of the angle between the direction of the sound wave and the normal to the ribbon. Since these directional properties are practically independent of frequency, they become useful in discriminating against undesired sounds and for obtaining a desired relation between the sounds from different sources and from reverberant sound in a studio. Its response[2] to reverberant or reflected sound is one-third that of a nondirective system, with the result that it can be used at a distance from a sound source of 1.7 times the distance of a nondirective type and still give the same results with respect to undesired reverberant sounds.

Because of the directional properties of the instrument, its sensitivity is at a maximum in directions in front and back perpendicular to the plane of the ribbon. With an input sound pressure of 10 dynes per sq cm the unit will normally deliver an effective output level of −54 dbm based on a reference level of 0.001 watt.

7. Moving-coil or Dynamic Microphone. This type of instrument (similar to Altec Lansing 633C) utilizes a light movable coil contained in a magnetic field to produce an emf which conforms with the sound waves impinging upon the dome-shaped diaphragm.

[1] Olson, H. F., *J. Soc. Motion Picture Engrs.*, **16**, 695, 1931; *J. Acoust. Soc. Am.*, **3**, 56, 1931.
[2] Olson, *loc. cit.*

The assembly is composed of a coil of fine aluminum ribbon edgewise wound and attached rigidly to a duralumin diaphragm of low mechanical stiffness which supports the coil in a radial magnetic field of a permanent magnet made from high-grade magnet steel. The diaphragm has a rigid dome-shaped center and a tangentially corrugated annulus. It has a high area/stiffness ratio. The diaphragm is cemented to a raised annulus on the outer pole piece. The outer and inner pole pieces are of soft iron and are welded directly to the magnet. The diaphragm is damped by an acoustic resistance which is supported below the coil by a brass ring, which in turn is held in place by rubber gaskets.

When the diaphragm vibrates in response to the sound waves impinging upon its surface, the coil vibrates in a like manner and cuts the magnetic lines of force. The emf generated in the coil is substantially proportional to the sound vibrations which cause the diaphragm movement.

The shape of the microphone housing and its size are such that the housing fits closely over the diaphragm and thus produces very little more diffractive effect than the diaphragm itself. To prevent resonance within the case an acoustic resistance baffle is provided to divide the space in two parts. A tube with its outlet at the back of the housing serves the double purpose of equalizing the inside and atmospheric pressures and of increasing the response of the instrument at low frequencies.

This microphone was designed to provide a uniform frequency response in all directions and has been termed a *nondirectional microphone*. For this reason the small cylindrical shape was selected as well as the method of mounting the diaphragm in a horizontal plane. A protective grid is provided over the diaphragm to control the resonance of the cavity in front of the diaphragm.

Wave-response calibration curves of this type of instrument indicate that the frequency characteristics are influenced to some extent at the higher frequencies by the angle of incidence from which the sound waves approach the diaphragm. Since the diaphragm is mounted horizontally, the instrument is entirely nondirectional with respect to the vertical axis.

The effective output level of the 633C microphone is −55 dbm as based on a sound input pressure of 10 dynes per sq cm and a reference level of 0.001 watt, and the frequency response is quite uniform over the range of 35 to 12,000 cps.

FIG. 9. Simplified cross-section view of the cardioid directional microphone.

8. Unidirectional Ribbon Microphone. In certain forms of studio technique it is desirable to eliminate the pickup of unwanted sound in the rear of the microphone, such as audience noise, room echo, etc. Here the unidirectional microphone is very useful.

The Altec Lansing 639A cardioid directional microphone utilizes a ribbon element of special design in combination with a compact pressure-type nondirectional element to secure a field response having a directional characteristic similar to a cardioid.

The pressure element is of the dynamic type having a dome-shaped diaphragm and constructed in some respects similarly to the dynamic microphone previously described. Commercial instruments of this general type have switches which enable the directional characteristics to be changed at will. The frequency-response characteristic of the type 639A microphone is quite satisfactory over a range of 40 to 10,000

cps. The output impedance is 30/50 ohms. It will deliver an effective output level of approximately −56 dbm as based on a sound input pressure of 10 dynes per sq cm and a reference level of 0.001 watt.

9. Polydirectional Microphone. The RCA type 77DX microphone combines the features of the velocity and pressure microphones to provide polydirectional characteristics by means of simple internal adjustments. This microphone is a high-fidelity instrument containing a simple ribbon suspended in an air gap formed by the pole pieces of a permanent magnet. The ribbon is divided into two individual sections, one of which is pressure-operated and the other velocity-operated.

The field response of the pressure-operated section is very nearly uniform in all directions and may be expressed as $E = E_0$, whereas the response of the velocity section is bidirectional and is equivalent to $E = E_0 \cos \theta$. Since the sensitivity of the nondirectional pressure section is made to equal the greatest sensitivity of the bidirectional or velocity section, the combined polar field-response characteristic of the two is equivalent to $E = E_0(1 + \cos \theta)$. In three-dimensional space, this is very nearly equivalent to a cardioid of revolution. The point of maximum sensitivity is directly to the front of the instrument, while directly to the rear the sensitivity approaches zero.

FIG. 10. Elements of the unidirectional ribbon microphone.

A uniform frequency response in all directions for the pressure-operated ribbon section is approached by allowing the operating face to be freely accessible to the atmosphere while the other side is terminated in an acoustic impedance very nearly equivalent to that of a very long pipe. Since a long pipe is too cumbersome for practical purposes, a short pipe of correct cross section, provided in coiled form and loaded through its length with absorbing material, such as tufts of felt, exhibits a suitable acoustic resistance over a frequency range covering all but the lowest frequencies.

The operating properties of the velocity-actuated ribbon section are the same as were described previously for the bidirectional velocity microphone. The ribbon is driven from its equilibrium position by a difference of pressure between the two sides, the pressure difference being due to the difference in phase between the two sides. The vibration of the ribbon caused by the sound waves impinging upon it causes an induced emf to be generated in the ribbon. The directional characteristics of the ribbon section are practically independent of frequency.

The polydirectional microphone differs from the conventional velocity microphone in that one entire side of the microphone ribbon is covered by a connector tube which is coupled to a damped pipe or labyrinth. Behind the ribbon in the connector tube an aperture in the form of an adjustable shutter is provided to permit various areas of opening between the ribbon and outside air. When the aperture is completely closed by the shutter, the microphone operates in the same manner as a pressure type. Under these conditions it has nondirectional characteristics. When the shutter is in the wide-open position, the ribbon is open to the air in front as well as to the rear and the microphone operates like the velocity type already mentioned. Under these conditions it has bidirectional characteristics. A common cardioid or unidirectional pattern is secured by an intermediate position of opening of the shutter. Directional patterns of the three types mentioned above are shown in Fig. 11. These are secured by settings of an adjustment screw on the rear of the microphone housing to marked positions.

To provide different amounts of l-f attenuation of the microphone response char-

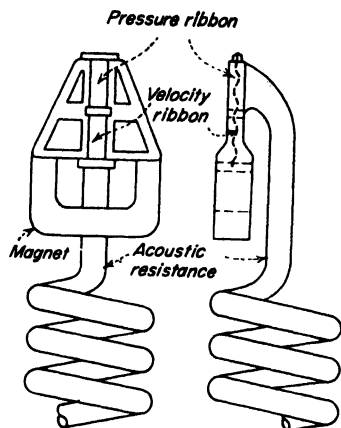
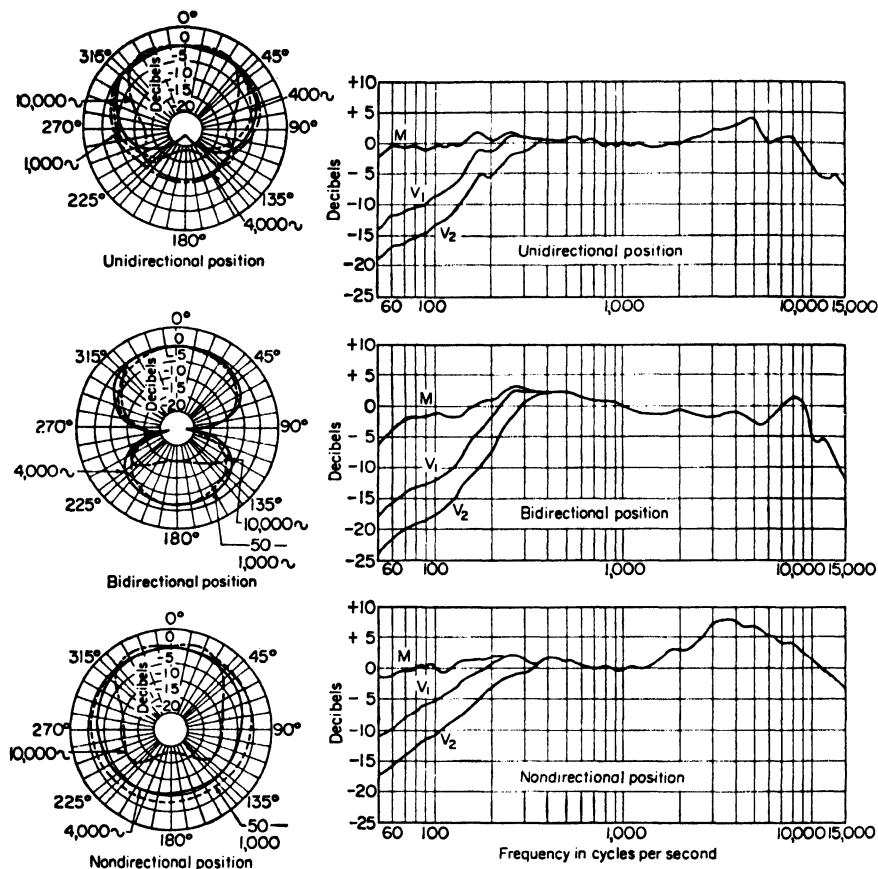



FIG. 11. Directional patterns and frequency response of 77DX polydirectional microphone.

acteristics under conditions of different directional settings of the shutter, a reactor is provided to shunt the output of the microphone transformer. A switch is provided within the lower shell to cut the reactor into the circuit as indicated by Fig. 12.

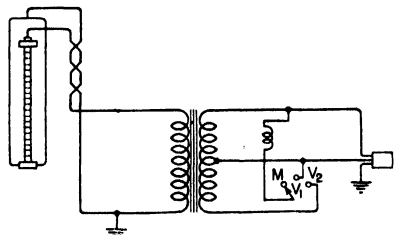



FIG. 12. Connections for type 77D microphone.

This microphone, by virtue of the polydirectional characteristics, provides a means of controlling the ratio of directional to reverberant sound as well as reducing unwanted sound such as audience noise. There are numerous applications of various directional patterns and the different response characteristics of this microphone to provide usefulness under various operating conditions. It is possible to secure any type of limaçon[1] directional characteristic by means of the shutter adjustment. Because of the variable directional characteristics with respect to the sound source at a point directly in front of it, the sensitivity is dependent upon

[1] Olson, H. F., Polydirectional Microphone, *Proc. IRE*, February, 1944, p. 77.

the directional setting. For bidirectional operation, the sensitivity is approximately equal to that of the velocity microphone previously described.

10. Crystal Microphone. This microphone utilizes the piezoelectric phenomenon produced in plates cut from piezoactive crystals. Thin plates cut from Rochelle-salt crystals are used almost entirely for the elements of crystal microphones. In comparison to other crystalline piezoelectric materials, such as quartz, Rochelle salt exhibits greater sensitivity for this purpose and it responds quite readily to mechanical vibrations.

Crystal microphones may be classified under two individual groups: (1) those utilizing multiple sound cells in free space and (2) those utilizing Bimorph[1] crystal elements each excited mechanically by an associated diaphragm. In the first of these types utilizing the Brush Development Company assembly, termed the *sound cell*, the elements are plates having dimensions 3/8 by 3/8 by 0.30 in. cut from Rochelle-salt crystals along axes in such a manner that their inherent characteristics tend to cause elongation or contraction when they are subject to an electric field provided by foil electrodes.

By cementing together two such piezoactive plates which have tendencies to act in opposition to each other when a voltage is applied, an assembly is produced with a motion analogous to the mechanical motion of bending a bimetallic thermostatic strip acted upon by variation of temperature. The assembly consists of two plate combinations mentioned above, separated by an airspace and held in position by a suitable mounting.

The cell is covered over with a membrane which serves as a pressure seal and protects the crystals from the outside atmosphere. When the cell is placed in a sound field, pressure acting normal to the outer surfaces of the plates tends to cause bending, with a result that an emf is generated between the foil electrodes. The two plate combinations are connected in parallel. The wave form of this emf conforms with that of sound waves. Because of the small physical dimensions of the plates, the frequency of mechanical resonance of the system is rather high, with the result that frequency response is quite uniform over a wide frequency range. Some models are quite uniformly sensitive up to 15,000 cps.

Commercial models contain series and series-parallel groups of these sound cells ranging from 2 to as many as 24. The sensitivity of a single sound cell is approximately −90 db, while a multicell microphone has a sensitivity as great as −68 db based on a zero reference level of 0.001 watt with a sound input pressure of 10 dynes per sq cm.

The output impedance (which is purely capacitative) of these instruments is quite high. This sometimes requires them to be operated directly into the grid of an amplifier tube having a grid leak of about 5 megohms. The small physical dimensions of a single cell make it practically nondirective. This property is also characteristic of multicell units. Figure 13 shows an *RC* amplifier suitable for use with such a microphone.

Fig. 13. Schematic diagram of amplifier for crystal microphone.

The diaphragm type utilizes a hermetically sealed Bimorph crystal supported at three points within the microphone housing. Projecting to the center of the specially treated fiber diaphragm is a small drive pin. This engages the remaining corner of the Bimorph crystal. Inasmuch as the Bimorph crystal is highly sensitive in converting fluctuating mechanical stresses, such as those caused by bending, into corresponding electrical fluctuations, the fluctuations in pressure created by the sound waves impinging upon the microphone diaphragm result in corresponding voltage fluctuations produced across the terminal ends of the Bimorph crystal plates.

[1] Brush Development Co. trade name.

Similar to most pressure-operated microphones, this instrument is semidirective. However, the smallness of the instrument assists in securing a rather uniform frequency response with direction. By placing the diaphragm facing the ceiling of the room, the instrument is essentially nondirectional in a plane through the diaphragm and parallel with the ceiling.

The output impedance of the crystal element is approximately 250,000 ohms. This does not permit long cables to be employed when high impedance connections are used. By means of a high-quality matching transformer this microphone can be satisfactorily operated into circuits exhibiting 50, 200, or 600 ohms impedance. The frequency response of this instrument is substantially flat from 100 to 5,000 cps. It has a variable control to allow manual adjustment of frequency-response characteristic while in operation. With an input sound pressure of 10 dynes per sq cm applied to the diaphragm the average crystal microphone of this type has an output level of approximately −48 dbm based on a zero reference level of 0.001 watt.

The crystal microphone shunted across the input resistor of the amplifier tube attenuates the higher frequencies of background noise such as those caused by thermal agitation. The properties of the crystals themselves are such that they are liable to damage, such as a change of frequency characteristics and output if the instrument is subjected to temperatures in excess of 120 to 125°F, particularly for periods of several hours.

11. Condenser Microphones. The condenser microphone utilizes the principle of mechanical variation of thickness of the air dielectric of a charged electrostatic capacitor as a medium to change acoustic energy into electrical energy of corresponding wave shapes. One form of this microphone consists essentially of an electric capacitor formed by a thin, tightly stretched duralumin diaphragm spaced approximately 0.002 in. and insulated from a flat brass disk called the *back plate.*

A polarizing potential difference is applied between the electrodes formed by the diaphragm and the back plate. The varying pressure upon the very thin diaphragm by the sound waves causes the electrostatic capacitance to vary by an amount in the order of 0.01 per cent of its normal value of 200 $\mu\mu$f.

The microphone has an aluminum alloy diaphragm 0.001 in. in thickness. The edges are clamped between threaded rings, the requisite stiffness being obtained by advancing the stretching ring until the desired resonant frequency is obtained. The space between the diaphragm and the back plate is hermetically sealed to prevent dust and moisture from entering and resulting in noise. The thin rubber auxiliary diaphragm, together with a small air-vent hole in the center of the back plate, is provided as an equalizing system for changes in atmospheric pressure.

On account of its inherent high-impedance characteristics, it is usual to incorporate an amplifier in the microphone housing to reduce to a minimum the length of the lead and the corresponding shunting capacitance between microphone and associated amplifier grid. A d-c polarizing voltage in excess of 180 volts has been used on some models.

Developments upon the early Wente[1] models by using duralumin as a substitute for steel as diaphragm material brought the sensitivity of modern instruments to about ten times that of early models.

Since this is a pressure-operated type of microphone, there are inherent irregularities in its characteristics from acoustic and mechanical phenomena. The microphone diaphragm is subject to certain resonance frequencies as well as the cavity. These tend to disturb the smoothness of the response characteristic. The sound waves striking and being reflected from the flat surface of the diaphragm cause pressure doubling, especially at high frequencies.

Below 500 cps this instrument is practically nondirective whereas at frequencies above 2,000 cps the directivity is noticeable. This directivity has a tendency to discriminate against h-f noise and reverberation, and, under certain conditions where

[1] Wente, E. C., *Phys. Rev.* 19, 498, 1922.

the studio does not accentuate the low frequencies, it has an advantage since the human ear responds more easily to background noise of higher frequencies than to lower frequencies.

The Western Electric 640A miniature condenser microphone unit[1] contains a diaphragm a fraction of an inch in diameter. The capacitance unit is mounted in one end of a tapered shell housing, of dimensions approximately 2½ in. in diameter and 7 in. long, which also contains the preamplifier. The weight of this microphone and preamplifier unit is 1¼ lb.

Published field-response curves indicate a rather uniform frequency-response characteristic from 40 to 10,000 cps with some dropping off of the higher frequencies as the angle is increased from an axis line normal to the diaphragm.

12. Parabolic Reflector Microphone. The use of a large concave reflecting surface mounted behind a microphone has been found to give the instrument pronounced

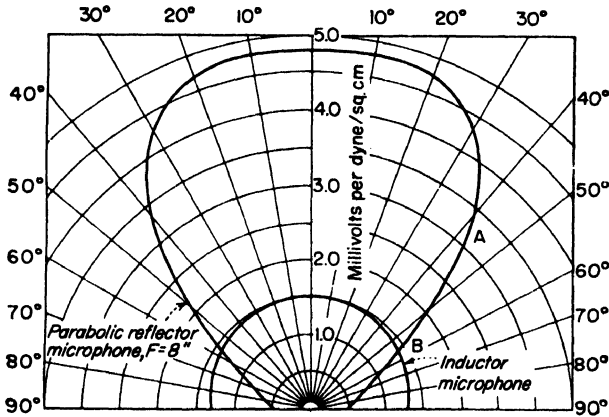



FIG. 14. Comparative axial response at 1,000 cps in millivolts per dyne per square centimeter. (*A*) Parabolic reflector; (*B*) dynamic-pressure-type microphone.

directional characteristics in the reception of sound waves. The system gets its name from the shape of the reflecting surface, a cross section of which contains a section of a parabola. By virtue of the microphone placement at the focus of the parabola of revolution or hollow paraboloid section, the sound waves striking the reflecting surface are concentrated upon the microphone diaphragm facing the inside of the paraboloid, resulting in increased sensitivity of the instrument in line with the axis inside of the paraboloid.

The use of the reflector, therefore, makes possible the placement of the instrument sufficiently far from the sound source so that it is practically equidistant from all the instruments or voices, with a result that the problem of securing proper balance and volume control is simplified. The directional characteristic makes it possible to swing the microphone and its reflector as one would a searchlight and in this manner follow the action on the stage of an auditorium or on the field of a sporting event. There is an increase in sensitivity along the line of axis of about 4:1, owing to the use of the parabolic reflector.

Since the reflector increases the sensitivity and makes it possible to locate the microphone at a greater distance from the source of sound, it is desirable that the output of the microphone should fall off rapidly if the sound originates at a point displaced more than 30 deg from the axis of the instrument; if this characteristic is obtained, reverberation and reflections in the studio or auditorium will have very little effect.

[1] Harrison, H. C., and P. B. Flanders, An Efficient Miniature Condenser Microphone System, *Bell System Tech. J.*, July, 1932, p. 451. Hopper, F. L., *J. Soc. Motion Picture Engrs.*, September, 1939, p. 278.

The h-f response may be increased as much as 15 db over the response at low frequencies by varying the position of the microphone in the reflector. However, in focusing the microphone, care must be taken to select the most useful frequency range, because at certain points of focus there is a tendency for sharp irregularities in the l-f response owing to cancellation between the directly received and reflected sound from the paraboloid reflector. In certain instances where the h-f absorption is considerable, the ability to accentuate the highs by refocusing proves very helpful.

Another distinct advantage of this directional microphone is its ability to disregard to quite an extent the acoustics of the room as it responds almost entirely to the sounds

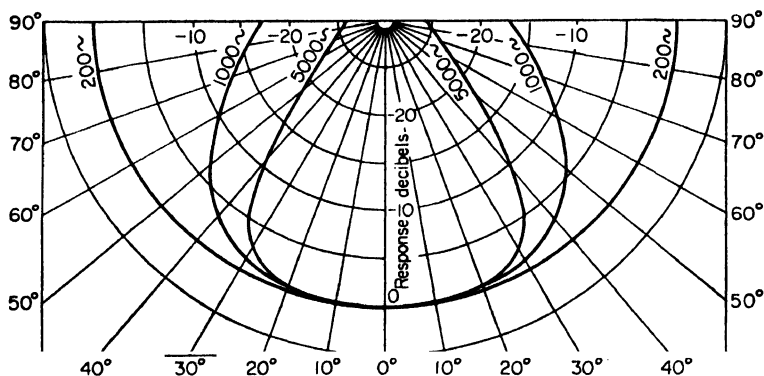



FIG. 15. Axial frequency response of parabolic microphone with a focal length of 8 in.

upon which it is directly focused. In some cases another microphone without a reflector has been used with the parabolic microphone so that it may be faded in at certain times to make the reproduction sound more realistic. The parabolic microphone has been used to pick up sound from a certain section of a large crowd or audience at a sports event or to pick up the voice of a single individual at a time in an audience.

13. Effective Output Level. When a microphone is connected to an unloaded input transformer its power output cannot be expressed in dbm because no appreciable power is delivered by the microphone. Therefore one must arrive at a figure which, when combined with the conventionally measured amplifier gain, will give the correct level for the combination, called the effective output level. It differs from the EIA standard rating G_M in the value of sound pressure and source impedance. The EIA rating computation is based on a source impedance of 150 ohms for all microphones having impedances between 75 and 300 ohms, and on a sound pressure of 0.0002 dyne per sq cm.

FIG. 16. Frequency response on axis of parabolic microphones having various focal lengths.

The effective-output-level calculation is based on the nominal microphone impedance and on a sound pressure of 10 dynes per sq cm.

The EIA standard defines the system rating G_M of a microphone as the ratio in decibels relative to 0.001 watt per 0.0002 dyne per sq cm of the maximum electric power available from the microphone to the square of the undisturbed sound field pressure in a plane progressive wave at the microphone position. Expressed

mathematically,

$$G_M = \left(20 \log_{10} \frac{E}{P} - 10 \log_{10} R_{MR}\right) - 50 \text{ db}$$

where E = open-circuit voltage of microphone
P = undisturbed sound field pressure
R_{MR} = microphone rating impedance (150-ohm)

Electrical reference level = 0.001 watt
Sound pressure = 0.0002 dyne per sq cm

For all practical purposes the output level of the microphone is obtained by adding to G_M the sound pressure level relative to 0.0002 dyne per sq cm. The sound pressure level of the program material can be measured with any of the several available sound-level meters. The exact relationship between G_M and the effective output level is illustrated below for the case of the type 44-BX velocity microphone connected for 250-ohm output impedance.

G_M = −146 db
+ 94 db sound pressure level for sound pressure of 10 dynes per sq cm
− 2 db correction for difference in source impedance 250/150 ohms

Effective output level −54 dbm

14. Hum Pickup Level. An arbitrary standard 60-cycle a-c field of 10^{-3} gauss has been established as a reference. It is fairly representative of fields measured at typical microphone locations in broadcast studios. The hum level is referred to 0.001 watt and is calculated in the same fashion as the effective output level, using as the output voltage the voltage produced by the standard field. A high-quality microphone has a hum pickup level in the order of −120 dbm or less.

15. Microphone Calibration and Testing. The sensitivity of a particular microphone is generally expressed as the open-circuit output voltage generated at the microphone terminals for a unit sound pressure against its active element. The intensity of the sound waves impinging upon the active element may be evaluated as a pressure or force. This is usually expressed in dynes per square centimeter.

The actual voltage generated by the microphone being very minute (a very small fraction of 1 volt for a sound pressure of 10 dynes per sq cm against its element), the sensitivity may be expressed in minus decibels below a standard reference level. When it is desired to correlate this value with the amplifier gain, one would have to assume the microphone to be loaded with a matching impedance. This would result in dropping the output voltage to one-half the open-circuit value or a corresponding 6 db decrease in output.

The sound pressure at a particular point where a standard microphone is set up is sometimes measured by the Rayleigh disk method. This instrument consists of a light circular mirror suspended by a fine quartz fiber at an angle of 45 deg to the axis of the tube through which the sound waves pass. The torque produced on the disk mirror by the sound field is measured by the deflected beam of light focused upon it. For small angles of deflection, the rotation of the disk is proportional to the sound intensity in the tube and consequently to the intensity in the undisturbed field. The actual value of torque may be determined by a torsion head which has a tendency to return the mirror back to its original position.

Where a sound chamber having suitable acoustic properties to prevent reverberation, at the lower frequencies especially, is not available, wave response calibrations are made in open air in a quiet atmosphere. From a standard microphone calibrated in this manner, other instruments may be compared to it for characteristics.

In determining the response characteristic of a diaphragm-type instrument such as a condenser microphone, use has sometimes been made of the thermophone method, the thermophone consisting of two strips of gold foil mounted on a plate and fitted into the recess in the front of the microphone, the recess being entirely enclosed and filled with hydrogen. A d.c. upon which is superimposed an a.c. is passed through the foil and causes fluctuations in the temperature of the foil and the gas immediately surrounding it. These

fluctuations in temperature cause changes in the pressure on the microphone diaphragm, and the magnitude of the pressure developed on the diaphragm can be computed from the constants of the system. Thermophone calibration is often referred to as a pressure calibration, since it depends entirely upon the actual pressure developed on the diaphragm and hence does not take into account any effects which may occur when the microphone is used for actual pickup purposes. The response obtained by placing the instrument in a sound field of constant pressure is termed a *field calibration.*

The effect of the diffusion of the sound field and the tendency for most acoustic materials to be more absorbent at high frequencies appear to cause the microphone actually to respond more closely to the field calibration than to the pressure calibration.

Previous to the use of any microphone in an actual broadcast or rehearsal, it is carefully tested by speaking into it and having another trained individual listen to the quality of the sound reproduced through a high-fidelity amplifying and loud-speaker system. As compared to the results secured from a standard microphone of known high quality, the condition of the microphone under test can be determined.

The outputs of two or more microphones when connected to a mixing circuit should have the same phase relation so that the output of one will add to that of another. Therefore, all microphones used for broadcasting are tested for phasing, and each one is connected to its connector plug so that proper polarity exists with respect to other microphones when it is plugged into a socket or jack.

Correct phasing of two or more microphones may be checked by first connecting one microphone to an associated amplifier input and setting the volume control knob to secure a given output reading on the volume indicator while talking into the microphone. The second microphone is then connected in parallel with the first and with the same volume control setting; both microphones are held close together and the same procedure of talking into them is repeated. A decrease in volume and impairment in quality from the previous signal would indicate that the polarity of one microphone was incorrect, and a reversal of connections on its plug to provide correct phasing with respect to the other microphone is required as indicated by an increase in volume with two microphones properly phased. Phasing of all microphones used should be checked in this manner so that the phasing of each agrees with all others.

STUDIO TECHNIQUE AND MICROPHONE PLACEMENT

16. Studio Problems. A problem of vital concern to a broadcasting system is that of providing favorable acoustic conditions within its studio or auditorium facilities so that the effects of reverberant sound from the walls of the enclosures may be kept within desirable proportions in comparison to the sound reaching the microphones directly from the source. Of even greater concern are the problems involving correct placement of microphones with respect to the sound sources within the enclosures, to assure faithful reproduction of each voice or musical instrument, their significant overtones, and a pleasant blending of the groups of voices or instruments.

It is, therefore, by virtue of the selection of a microphone which will faithfully transmit all the actual sounds that occur within its range as well as the correct placement of it within a studio or auditorium having suitable acoustic characteristics that high-quality programs can be produced. Under optimum conditions of reproduction a broadcast listener would hear the same acoustic naturalness of the program from his loud-speaker as he would if he were to be transported to a favorable spot in the studio.

The major considerations involved in proper studio design such as soundproofing, ventilation, optimum dimensions, and suitable acoustical treatment of the walls have been given.[1] At present we shall be concerned only with the problems of microphone placement, assuming that favorable studio and auditorium conditions exist. Normally, these considerations would be as follows: adequate soundproofing that would

[1] Hanson and Morris, *loc. cit.* Sivian, L. J., *Bell System Tech. J.*, **10**, 108, 1931. Morris, R. M., and G. M. Nixon, Broadcast Studio Design, *RCA Rev.*, October, 1936. Nixon, G. M., *RCA Rev.*, January, 1942, p. 259. Content, E. J., and L. Green, Acoustic Design and Treatment for Speech Broadcast Studios, *Proc. IRE*, February, 1944, p. 72.

prevent undesired extraneous noises from entering a given enclosure, and suitable acoustical treatment of the walls and floor to provide equal absorption over a wide frequency range and give the enclosure in itself a uniform frequency characteristic. It is of considerable importance that the frequency characteristic of the studio or enclosure be considered for high-quality transmission because this characteristic is actually superimposed upon that of the microphone under conditions where the reverberant sound received by the microphone is appreciable as compared with that received directly from the source.

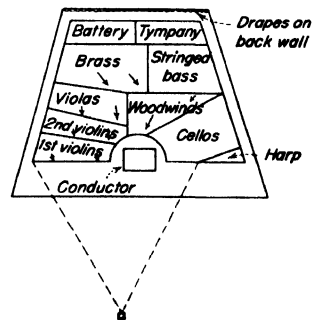
17. Single vs. Multiple Microphone Usage. During the first years of broadcasting, it was a usual procedure to use more than one microphone to pick up a program, especially under conditions where the broadcasting group was rather large. This was necessary on account of rather low microphone sensitivity and the inherently high noise level of the microphones used during that period requiring a placement of those instruments sufficiently close to the sound sources to overcome the inherent background noise. The combination of more than one microphone for making a pickup has a disadvantage in that the outputs from the several microphones used were not in proper phase relation with respect to the sound sources. This resulted in considerable distortion when the microphone outputs were combined and fed into a common amplifier.

Improvement in microphones to secure higher sensitivity as compared to inherent instrument noise level has resulted in the possibility of using one microphone at a time. Each microphone is located at a sufficient distance from the sound sources so that more than one microphone is not necessary to obtain a good acoustic balance from a particular group. The practice of using more than one microphone at a time has therefore been discouraged whenever possible because of the phase distortion in the sound field resulting therefrom. Under conditions where more than one microphone is used at a time, each should be properly phased or polarized with respect to others as mentioned above.

18. Microphone Placement. For diaphragm types of microphones, the directional characteristics at the higher audio frequencies may require the placement of the broadcasting group located in front of the diaphragm to be contained in an area formed by an angle of 30 deg on either side of the microphone axis.

The frequency characteristics of any diaphragm type of microphone are dependent upon the relative positions of the microphone and the sources of sound in front of the diaphragm. When the sounds approach at right angles to the plane of the microphone diaphragm, a uniform response over the desired range might be obtained. But, if the sounds approach from any other point, it will be found in general that the response will fall off with frequency. This characteristic indicates that response varies with the angular displacement of the sound source from the microphone axis. There is a high loss at the higher frequencies for high angular displacements from a point in front of the microphone. Since most musical instruments depend for their quality or timbre upon the presence of overtones, it is obvious that, if these overtones are discriminated against, the quality will be changed materially. Considering this loss in the higher frequencies with angular displacement and using a single microphone of the diaphragm type, all the musical instruments of a group should usually be kept within an angle of 30 deg either side of the microphone axis. Present-day microphones of the diaphragm type have uniform frequency response over much wider angles and permit arrangements of the broadcasting groups as shown in Fig. 17 within an angle of 45 deg on either side of the microphone axis.

Fig. 17. Microphone setup for large orchestra.

An individual source of sound such as a speaker, announcer, or musical instrument

should not be placed closer to the microphone than 1 ft. Greater distances are determined by the volume range of the voice or instrument and the relative volume desired with respect to the accompanying instruments.

One must consider that in different selections and different arrangements of the same selection the relative importance of the particular instruments may be changed considerably. Where desired prominence cannot be given to a particular group at a certain time using a single microphone, it may be necessary to fade-in another located near the group to be emphasized. A number of microphones can in this way be used, in the particular interval desired, to emphasize the particular musical instrument or

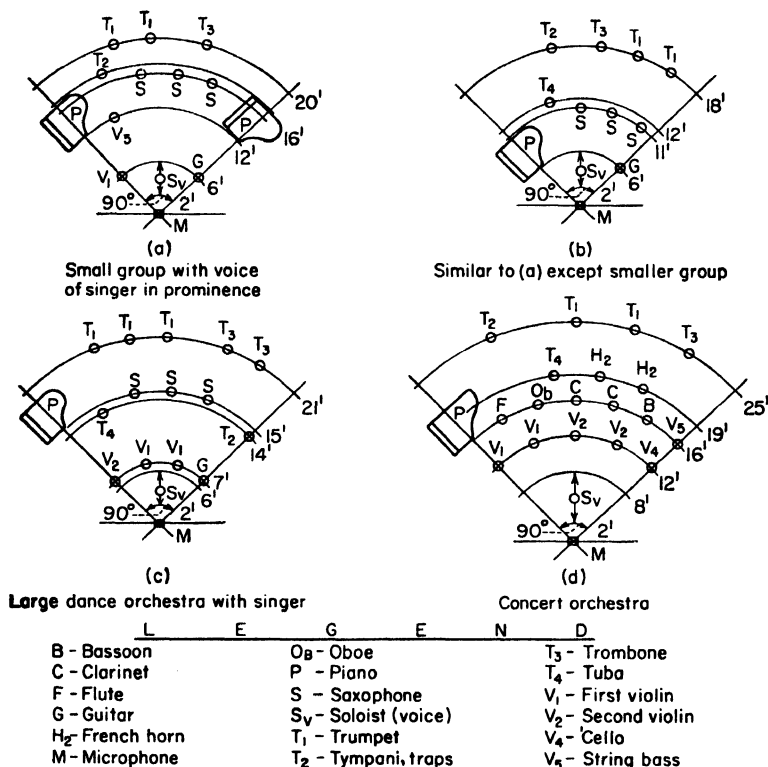



FIG. 18. Orchestra arrangements for use with a single diaphragm-type microphone.

instrumental group, the soloist, or the announcer. The control of individual microphone circuits for this arrangement is performed in the control booth by suitable mixing and switching devices. Extensive rehearsing accompanied by listening tests at a remote point are generally required to secure the desired balance for a particular setup.

In general, the volume range of certain instruments adjacent to one another permits their alliance into natural groups, each instrument of a group being placed approximately equidistant from the microphone. One group may contain violins, violas, and cellos; a second group, the piano, harp, flutes, and clarinets; a third group, the oboes, bassoons, and French horns; a fourth group, the string bass, tuba, timpani, and traps; a fifth group, the trombones and trumpets. In dance orchestras the guitar is usually placed in the first group, the saxophone in the third group, and the banjo with the fourth group.

There are many factors involved in securing the proper placement of vocal sound

sources or musical instruments before a microphone, particularly before a pressure or diaphragm type. While certain rules have been set up, they may serve only as a guide. Most satisfactory results are obtained by a combined study of the instruments as well as an actual setup of them before a microphone in a given enclosure. The result of actual listening tests by means of a high-fidelity speaker and monitoring system performed by one who has a trained ear for music or sound naturalness is a final check upon the proper placement.

19. Typical Studio Arrangement. A typical setup of a large orchestra before a directional microphone is shown in Fig. 17. The instruments are placed to obtain the desired balance for theater or auditorium work and to obtain the proper harmonic balance allowing for the microphone directional characteristics. The microphone is usually placed high above the floor, sometimes near the ceiling of the auditorium.

Present-day microphones, such as the unidirectional and cardioid types, are used to advantage since their response in the rear is very small. The string instruments in this setup, being the least powerful ones, are concentrated in the foreground of the group. The wood winds are next in line, followed in the background by the powerful brass and percussion instruments. In the arrangement the string tone of the orchestra is given a favorable position to produce a softness to the music which will not be overpowered acoustically by the heavy brasses and percussion instruments.

FIG. 19. Microphone arrangement for dance orchestra.

Figure 18*a*, *b*, *c*, and *d* shows various arrangements of instruments and voices before the pressure or diaphragm type of microphone. The directivity vs. frequency-response characteristic of this particular type permits the placement of the musical instruments within an area contained by an angle of 45 deg on either side of the microphone axis. In using this type of instrument the source of sound, speaker, announcer, or musical instrument should not be placed closer than 1 ft from the face of the microphone.

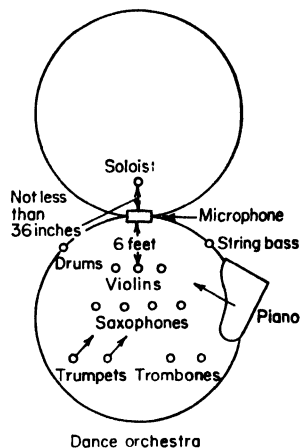
The bidirectional characteristics of the velocity microphone are advantageous in that the performers can be distributed on both sides of the instrument in a manner shown in Fig. 19. The uniform frequency-response characteristic of the instrument with directivity is an advantage in that the intensity of some instruments may be decreased without discriminating against their higher frequencies, simply by moving them at a larger angle with respect to the microphone axis.

An orchestral arrangement involving the use of a velocity microphone[1] is shown in Fig. 20.

The orchestral group in this arrangement was conveniently located on one face of the instrument. To prevent reflection from a wall directly in back of the microphone, the instrument is tilted at an angle of approximately 30 deg toward the orchestra. An exceedingly well-balanced pickup has been accomplished by this method.

20. Volume Controls or Faders. Volume controls or faders used in high-quality broadcasting circuits should have frequency characteristics which are uniform between 30 and 15,000 cps to prevent them from causing frequency distortion. Also essential is a very low noise level. This is normally -150 db or better. Proper shielding for protection against dust and dirt is necessary to maintain a low noise level, as well as to act as a shield against any stray r-f electromagnetic fields.

In Figs. 21 to 27 are shown various types of attenuating structures used in broad-

[1] LaPrade, Ernest, The Technique of Broadcasting Instrumental Groups, address at North Central Music Educators Conference Indianapolis, March, 1935. See *Proc. Music Educators Nat. Conf.*, 1935.

casting technique. The type shown in Fig. 21 is frequently used as a microphone fader and is commonly known as the LT structure. When used in multiple such as for mixing several microphone outputs, as in Fig. 22, sufficient resistance is inserted

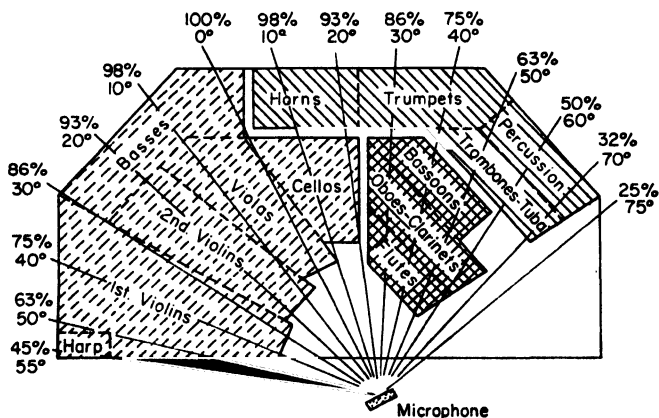



FIG. 20. Velocity-microphone setup for large orchestra group.

in one output lead from each attenuator to maintain correct circuit matching. The bridged-T structure shown in Fig. 23 is used extensively for the same purposes.

The ladder attenuators maintain an impedance that remains practically constant in both directions through the middle of the attenuation range. Important features of this type of attenuator are its simplicity of design requiring fewer contacts and switches. The minimum attenuation setting of a ladder pad normally corresponds to its insertion loss, which amounts to approximately 2.5 db. Where an attenuation range is required extending from zero upward, the H or T structures are used. They are usually constructed with a minimum attenuation setting of zero.

FIG. 21. LT attenuator.

The T and balanced-H structures maintain a constant impedance in both directions when properly terminated. The balanced-H and ladder structures are used where

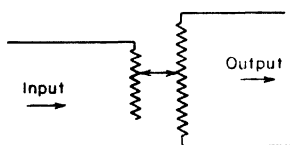
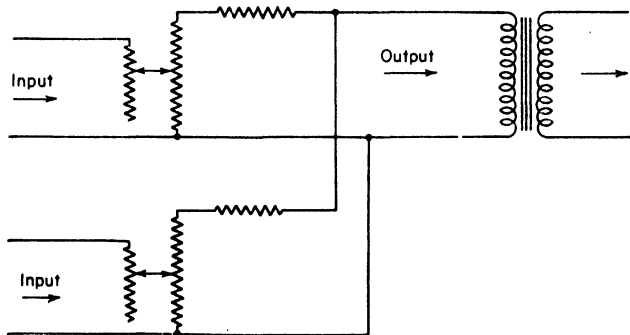



FIG. 22. Multiple-type LT attenuator.

the transmission circuits must be balanced to ground for the purpose of avoiding cross talk and noise that are present in unbalanced circuits. They are frequently used in broadcasting circuits as master gain controls. Figure 28 shows a high-imped-

ance voltage divider usually in the form of a gain control in the input circuit of a vacuum tube. This is a common type of gain control used on speech amplifier units.

Microphone fading is usually accomplished at high level or at the outputs of the preamplifiers and in a control booth beyond where programs originate in studios. For field pickups fading is sometimes accomplished directly at the outputs of the microphones. This, of course, requires attenuators of very low noise level. Microphones of the moving-coil dynamic and the velocity-ribbon types have constant low-impedance output over a wide frequency range and for this reason can be faded directly at their outputs.

Fig. 23. Attenuator of the bridged-T type.

The curves in Fig. 29 give resistance values of the branches of a T or H pad suitable for a channel having an impedance of 200, 500, or 600 ohms, the range of attenuation being between 2 and

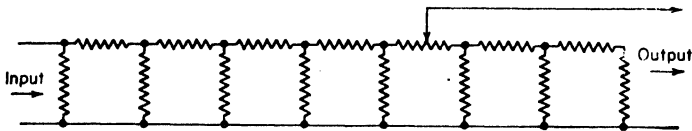



Fig. 24. Single-ladder attenuator.

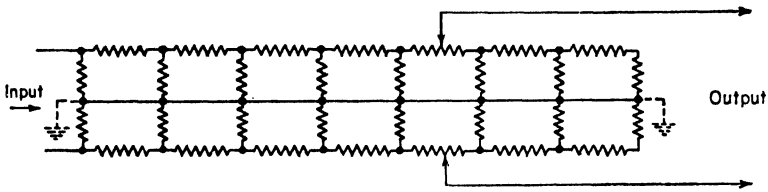



Fig. 25. Balanced-ladder attenuator.

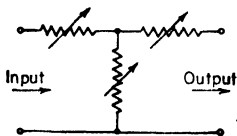



Fig. 26. Type T attenuator.

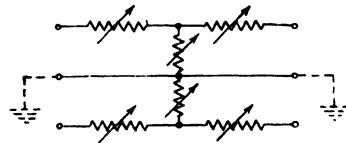



Fig. 27. Balanced-H attenuator.

30 db. Similar curves for other impedances may be determined from the formulas published previously.[1]

Fig. 28. High-impedance voltage divider.

21. Volume Indicators. The volume level of an audio signal at any particular point in a broadcasting system is normally measured by means of a standardized instrument called the *volume indicator*. The components of the complete instrument consist essentially of a sensitive high-resistance voltmeter of the copper-oxide type, an associated T attenuator for extending the range of the meter to higher readings, and a variable resistor accessory to the attenuator provided for calibrating the instrument. The instrument scale is marked

[1] Johnson, K. S., "Transmission Circuits for Telephone Communication," Van Nostrand; and Lanterman, W. F., The Design of Attenuating Networks, *Electronics*, February, 1931.

in a logarithmic fashion, and superimposed upon this is an associated percentage scale. Two scales are provided, the A type tending to emphasize the vu readings and the B type in which the percentage readings are more prominent.

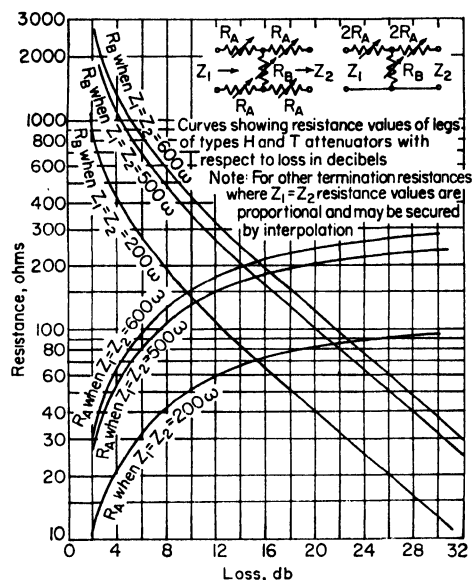



FIG. 29. Chart for H and T attenuator design.

While an oscillograph placed across the circuits at a particular point in the system would give a true picture of the rather complex wave shapes present from program signals, it would be a rather cumbersome and expensive method of indicating the characteristics of the signal, although it could be used if properly calibrated against a standard. However, through coordination between the broadcasting systems and the telephone company whose facilities were also involved, there was developed[1] a standard instrument of the indicating-needle type having characteristics most suitable for the purpose of indicating signal volume.

The standard volume indicator (Fig. 30*a* and *b*) utilizes a d-c instrument with a noncorrosive full-wave copper-oxide rectifier mounted in its case. Arranged for bridging, as in Fig. 30*a*, across a line, it has an impedance of about 7,500 ohms measured with sinusoidal voltage. Of this impedance 3,900 ohms is in the meter and about 3,600 ohms is external for the purpose of securing required dynamic characteristics.

The dynamic characteristics are such that if a 1,000-cps sine-wave voltage of an amplitude to give a steady reading of 100 on the voltage scale is suddenly applied, the

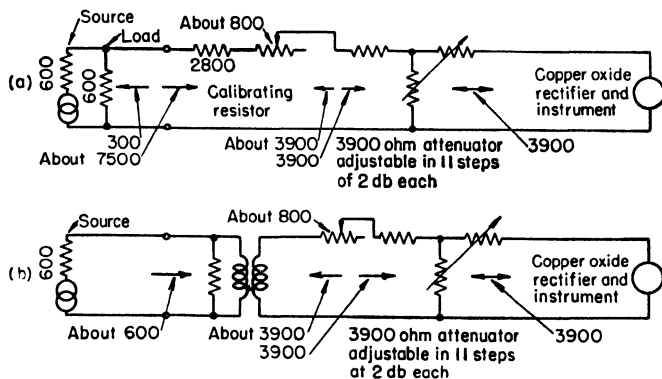



FIG. 30. Standard volume-indicator circuit. (*a*) For bridging across a line; (*b*) low-impedance arrangement as for line termination.

pointer will reach 99 in 0.3 sec and then overswing the 100 point by at least 1.0 and not more than 1.5 per cent. The frequency response of the instrument is very good,

[1] Chinn, Gannett, and Morris, *loc. cit.* See also Chap. 13.

as is indicated by the fact that it does not depart from its 1,000-cps reading by more than 0.5 db between 25 and 16,000 cps.

The standard volume indicator is calibrated to read 0 vu when it is connected to a 600-ohm resistance in which is flowing 1 mw of sine-wave power at 1,000 cps or *n* vu when the calibrating power is *n* db above 1 mw. However, owing to limitations in the present art, it has not been found practical to make an instrument of sufficient sensitivity to be calibrated to read 0 vu across 600 ohms with 1 mw, and therefore the instrument is normally calibrated by the application of 1.228 volts rms (4 db above 1 mw in 600 ohms) to the instrument in series with the proper external resistance to cause a deflection to the 0-vu or 100-scale point. The instrument therefore has sufficient sensitivity to be read at its normal 0-vu point on a volume level of +4 vu, which

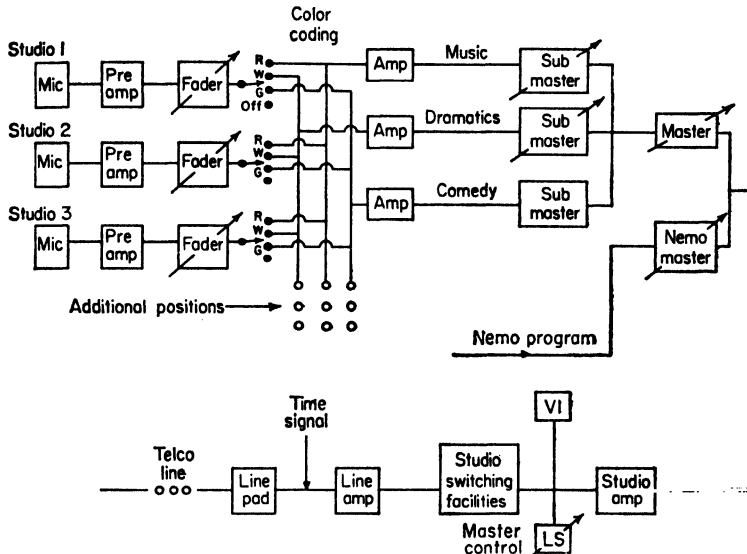



FIG. 31. Equipment and facilities required for a single program channel.

is the minimum setting of the attenuator and volume indicator; for this reason the 1-mw calibration is correct.

For routine checking of the calibration of volume indicators, a "reference" instrument properly calibrated may be used in a simple comparison method.

To the terminals of a source of a-c voltage of adjustable output, the reference volume indicator and volume indicator to be calibrated are connected in parallel. The attenuators of both indicators should be set at +4 vu. The applied voltage is then adjusted until the reference-volume-indicator pointer is at the 0 vu or 100 mark. If the pointer of the volume indicator being checked is not then on the 100 mark, its calibration resistor should be adjusted until it reads the same as the reference volume indicator.

Since the standard volume indicator has been developed and standardized as a method of checking the volume of signals of complex wave shapes, it has associated with it the term vu. This term has been restricted to its intended use; hence, whenever a volume level reading is encountered expressed in so many plus or minus vu, it will be understood that the reading was made with an instrument having the characteristics of this standard instrument and expressed with respect to the reference level. Certain types of volume indicators, even when recalibrated to a 1-mw basis, will not give indications corresponding to those of the vu instrument on all types of program

waves owing to the particular characteristics of the instrument. The zero reference level for communication purposes is usually expressed as zero dbm. This level is based on a 1-mw signal of single-sine-wave frequency as measured in a circuit having a standard 600-ohm resistance termination. See also Chap. 13, Sec. 35.

22. Speech-input Amplifiers. These amplifiers are also termed *preamplifiers* or *microphone, line,* and *program amplifiers.* They comprise the apparatus necessary to increase the electrical energy output of the microphone or transcription reproduction to a sufficient level to permit its transfer by means of wire lines to the broadcast transmitter. The normal energy level of programs entering the wire lines or program loops is approximately +8 vu (+14 vu delivered from the line amplifier with a 6-db isolating pad). A schematic of the equipment shown in Fig. 31 shows positions of the microphone controls, volume indicators, monitoring amplifiers, and relay-switching systems.

Speech-input equipment is designed to have a substantially uniform response from about 30 to 15,000 cps and above. The maximum gain of such a three-stage amplifier from input to output is approximately 75 db. Input impedances are 62.5, 250, and 600 ohms, and the output impedances are 250, 500, and 600 ohms.

PROGRAM RECORDING FACILITIES

The essential parts of a large broadcasting system usually include the facilities for recording programs for the following reasons:

1. To have an accurate record or log of the program material actually broadcast from a station. This is known as *reference recording.*
2. To secure a record of a studio or special-events program at some distance over wire lines and thus be able to reproduce the program at a time most convenient for an audience which may be in a time zone a number of hours different from that in which the event takes place.
3. For production of recordings to use at small stations where wire-line facilities are not available.
4. The recording of an audition of a person or group of persons qualifying for a program part.
5. Production of sound effects such as crowd noise, for convenient use and introduction into a particular program.

23. Recording Equipment. The essential equipment required for producing high-fidelity recordings on magnetic tape or disk records consists of the following: For bridging a program bus by means of a multiple point switch there is a limiting amplifier of the type similar to that described in Sec. 48. It is the function of this limiting device to prevent overcutting of the record on high peaks. Following this are duplicate recording channels, each a program amplifier having linear amplitude characteristics and a uniform frequency response over a wide range. This amplifier normally has audio power-handling capabilities up to +46 vu so that high audio peaks are not distorted before reaching the cutter head. There is a standard volume indicator across the line following this equalizer since the cutting head is placed after the equalizer, the output of which drives the cutter head.

24. Sound on Disk Recording.[1] The direct method of disk recording utilizes aluminum disks usually 16 in. in diameter and 0.050 to 0.060 in. thick, coated with a cellulose nitrate compound (usually miscalled "acetate"). The cellulose nitrate coating is used as the medium for recording sound modulations. These disks are suitable for immediate playback.

For indirect recording it is the usual procedure to use soft wax records approximately 17 in. in diameter and from 1 to 2 in. thick. These records are later processed to produce a hard record approximately 16 in. in diameter and ¼ in. thick.

The sound record is cut in the highly polished surface of the wax disk by means of an electromechanical recorder. The technique of cutting wax records is similar to making standard electric phonograph records. The standard speed for common phono-

[1] See Chap. 19, Secs. 111 *et seq.*

graph records is 78, 45, and 33⅓ rpm while for broadcasting records it is usually 33⅓ rpm. This speed with a 16-in. disk gives a playing time from 10 to 15 min.

Both types of records are usually cut with the spiral proceeding from the outside edge of the record toward the center, similar to making standard electric phonograph records.

25. Variation of Frequency Response on Disks. In recording on a disk revolving at a constant angular velocity, the cutter stylus is placed near the outside edge of the record, and, as the engraved spiral of the sound track progresses toward the center of the disk, the velocity with which the stylus travels on the disk is decreased. This velocity is directly proportional to the radius between the center of the disk and the position of the stylus. Now, for most satisfactory reproduction of the higher frequencies, the stylus must travel with sufficient velocity over the disk to provide sufficient space in the groove to permit satisfactory engraving of the h-f pulsations of very short duration. Therefore there is a tendency for more satisfactory engraving of the higher frequencies near the outer edge of the disk than toward the center. In other words, for a given cutter engraving on a disk of constant rotational speed, the frequency response one may reproduce from a disk is more satisfactory near the outer edge of the disk and is less satisfactory, especially at high frequencies, as the stylus moves toward the center. For this reason, if high-fidelity results are to be obtained, the groove containing sound modulation should not be closer to the center of the disk record than 5 in. for 78-rpm recording and 8 in. for 33⅓-rpm recording. Medium to good results are obtained with the groove containing the sound modulation at a radius on the disk of not less than 2½ in. for 78-rpm and 4 in. for 33⅓-rpm recording. For a given playing time it is sometimes possible to keep this minimum radius, cutting more grooves per inch, sometimes as many as 160 in lateral and vertical disk engraving.

26. Lateral and Vertical Disk Engraving. In the lateral system the groove depth is kept constant, and the engraving stylus moves in a horizontal fashion to produce undulations in the sides of the groove. The groove spacing therefore must be sufficient to prevent the stylus from cutting into adjacent grooves at the low frequencies.

The vertical system utilizes an engraving stylus moving in a vertical direction. The groove depth varies with the mechanical modulation, whereas the groove width is kept more or less constant, with the result that the groove spacing can be kept closer with a correspondingly greater duration of playing time.

In Fig. 32 are illustrated waves produced in disk-record grooves under conditions of "constant-amplitude" and "constant-velocity" recording. The wave marked 1 illustrates constant-amplitude engraving produced by a constant sound level regardless of frequency at the cutter. In this case the recorded amplitude is the same for all frequencies.

The constant-velocity system utilizes constant vibrational velocity of the stylus in the record groove under influence of the cutter head. In this case the amplitude of the wave is inversely proportional to the frequency. The wave marked 2 illustrates undulations in the record groove produced by constant-velocity recording, producing an increase in amplitude with a decrease in frequency for a constant sound-level input (assuming that the entire system from the microphone to the cutter head has a uniform frequency characteristic). In this instance, where amplitude $= v/kf$, the amplitude of a wave frequency f of 100 cps would be one-half that of 50 cps for a constant velocity v. At the lowest frequencies, therefore, the amplitudes would be excessive if sufficient amplitudes of the higher frequencies are to be produced in the record groove. Since the groove spacing would have to be made considerable to avoid "groove crossover" or "echo" effects in adjacent grooves, due to excessive amplitudes at the lower frequencies, it is customary to cut records at constant amplitude for frequencies below some point between 350 and 800 cps and at constant velocity for frequencies above this point. This is illustrated as wave 3, a solid line. The transition frequency between constant-amplitude and constant-velocity recording, normally some point selected between 350 and 1,000 cps, is called the *turnover point*.

To produce a constant-amplitude cutting characteristic up to the turnover point

and a constant-velocity cutting beyond, it is necessary to utilize corrective equalizers depending upon the particular type of cutter head used. In some cases the response characteristic of the electrodynamic cutter head itself is a contributing factor in the production of the constant-amplitude and constant-velocity range as well as the turnover point. This is illustrated in Fig. 33.

Commercially, direct recording has become of great importance because of its advantages of immediate playback and cheapness when producing disks in small quantities. While the nitrocellulose coating is essentially softer than the pressed

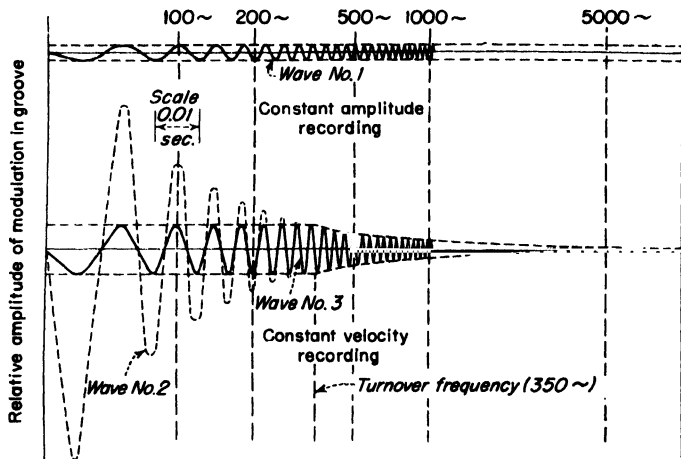



FIG. 32. Characteristic of waves produced with the constant-amplitude and the constant-velocity systems of recording.

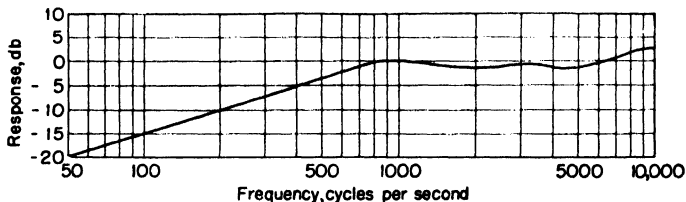



FIG. 33. Frequency response of cutter head based on optical measurements of the stylus-tip motion for constant input.

records manufactured by the electroplated soft wax process, as many playbacks as 100 may be secured from a nitrocellulose disk with a well-designed lightweight reproducer. A substantially flat frequency response may be recorded on, and reproduced from, these disks over a range of between 50 and 10,000 cps and higher near the outside portion of the disks. It is good practice where very high-fidelity reproduction is required on 33⅓-rpm disks to use the outside portion of the disk to compensate for the loss of the higher frequencies in reproducing as the pickup moves toward the center of the disk or to divide time into two or more disks, thus permitting reasonably high linear recording velocity of the cutter stylus. A volume range of approximately 55 db has been obtained from nitrocellulose disks using the lateral system of recording and reproducing. With satisfactory operating conditions over-all distortion of the combined recording and playback operations is less than 5 per cent. This over-all distortion is also a function of engraving velocity, decreasing as the velocity is increased, also decreasing with a decrease in engraved depth.

Flutter is a term used to describe vertical modulation produced in the recording groove due to the bounding of the cutter head at a frequency of approximately 30 cps. It is normally caused by mechanical response of the recording head and its associated supporting-arm mechanism under excitation from building noise and other l-f rumble. In observing reflections from record grooves created by a single source of light, the effects of flutter can be noticed in the form of spokes or long spiral patterns extending from the inside over the recorded surface to the outside. Under a microscope this vertical modulation may be seen as a varying width of the cut groove. Manufacturers supply stabilizers which assist in the elimination of flutter.

When recording on nitrocellulose disks, an air-suction nozzle is provided near the cutter to remove shavings or shreds so that they will not interfere with the engraving process and also to provide for safe disposal of this highly inflammable material. Care must be taken to avoid dust, fingerprints, or grit from entering the engraved surfaces of the disk. Otherwise there is a tendency for increased noise. It is customary to engrave 120 grooves per inch on these disks, although 96 and 112 and as high as 160 grooves per inch have been used. This number is fixed by the lead screw of the recording machine. The groove depth engraved on this type of disk is normally about 0.0015 to 0.002 in. Commercially, it has been possible to secure recordings of this type having a noise level 50 to 60 db below the maximum modulated signal, although the average record has only a 35- to 40-db spread between noise and modulated signal. By the method explained below for processing soft wax from which pressings are made of a hard material, nitrocellulose disks may be similarly processed for the purpose of making a large number of pressings.

The indirect recording method requires considerably more equipment and time to manufacture the pressed disks than the direct method described above. However, for mass production, pressings can be made considerably more cheaply than single records by the direct process.

27. Necessary Equipment. Equipment necessary for wax disk recording consists essentially of a machine lathe especially designed to turn the wax record clockwise at a uniform speed, which is 33⅓ rpm for broadcasting work. The carriage of the lathe is driven with a lead screw carefully machined to move the recorder holder at a predetermined rate while cutting the wax record. The lead screw is driven through a gear train which regulates the number of grooves cut per inch, usually 86, 92, 98, 112, or 120. A recorder holder provides the necessary support for the electrical recorder.

A horizontal turntable, driven through a vertical shaft, is provided for supporting the wax record. The vibration of the driving motor is eliminated on different lathes by various methods.

28. Disk Records. The grooves of a disk record are ordinarily spaced 92 to 160 per inch. For 92 grooves per inch this allows about 0.011 in. from center to center of the groove, of which 0.006 in. is the width of the groove itself. The maximum lateral motion of the stylus is thus limited to about 0.0025 in. on either side. Generally, 0.002 in. should not be exceeded. Cutters usually used are designed as constant-velocity devices. In practice such cutters have this characteristic only above 300 cycles or higher. Below this point the amplitude is independent of frequency. If the maximum amplitude for a 300-cps wave is equal to 0.002 in. on either side of the center, then a 1,500-cps amplitude for the same electrical input level would be 0.0004 in.

The shape of the groove varies somewhat in commercial practice, but it is approximately 0.006 in. wide and 0.0025 in. deep. The pitch of the groove is generally 0.010 to 0.011 in., leaving a space between grooves of about 0.004 in. With only this space available, the maximum safe amplitude is something less than 0.0025 in., if the walls of the groove are not to be cut too thin.

Cutting stylus consists of a sapphire, synthetic ruby, or other hard point fastened to the lower end of the stylus arm. One end of the sapphire has a rounded point about 0.002 in. radius and standard cutting angles of approximately 70 or 90 deg for the sides.

The advance ball is a small cylindrical sapphire, ground spherically at one end and

held in an adjustable mounting attachment to the recorder. This ball supports the weight of the recorder, and the arm, being adjustable, permits regulation of the depth of the groove on the wax.

Playback reproducer is provided to permit playing back the wax record immediately after it is cut for rehearsal work and test. This usually renders the wax unsuitable for processing, and for this reason two wax records are usually provided for each recording channel, one of which can thus be used for playback and the other for processing. The pressure of the needle on the wax is generally adjusted to between 15 and 20 g.

A needle provided for playback from the soft wax is designed differently from the ordinary needle used for the finished hard record. The Western Electric type has a point of 0.003 in. radius. The needle is constructed on a mandrel, ground to a smooth finish, and the point given a chromium plate to improve wearing quality.

Checking Speed. The periphery of the turntable is usually divided with vertical lines, so that a neon lamp, operating from a 60-cps source, may be used as a stroboscope to observe the turntable motion. The lines on a standard turntable are usually arranged so that with 60 cps on the lamp, as the turntable rotates at exactly 33⅓ rpm, the lines will appear to be stationary. If faster than 33⅓ rpm, the lines will advance slowly, and, if slower than 33⅓ rpm, the reverse will be the case. This check of the speed is usually made with the wax record on the turntable.

Checking the Damping Action. A method of checking the instantaneous constant speed may also be used to check correct damping of the turntable. With the turntable rotating at normal speed, the oscillator for supplying 60-cps source to the neon lamp may be adjusted until the vertical lines appear stationary. If the disk is now touched lightly by hand, the line or spot observed will appear to shift its position owing to momentary load. As soon as the hand is removed, the line or spot observed should come back to its original position. Observing the movement will determine whether the turntable has insufficient damping or too much damping.

Determining the Starting Point. Disk records for radio broadcasting use are usually cut in clockwise rotation from the outside in, similar to ordinary phonograph disk records. To obtain a definite starting point for the records when in use, the first groove is spaced an appreciable distance from the rest of the cut. This is obtained by a coarse speed cam actuating the lead screw at the start of recording. As the lead screw makes its first complete revolution, it moves the recorder under the influence of the cam until the recorder is in its normal cutting position.

29. Wax-suction Equipment. This equipment is provided to furnish a means of removing the shavings from the wax record during recording. The suction tube is so placed that the shavings thrown off by the stylus are carried away from the face of the wax. A central suction system is usually provided in studios having several recording channels. This usually consists of a turbine suction pump with pipe lines leading from a central suction point to a separator tank placed in each recording room. In some smaller installations, an individual bell jar, with a small suction motor, is used for each recording machine.

30. Wax Preparation. Two types of waxes are generally used in sound recording, those having a working temperature of 75°F and those with a working temperature about 90°F. Matthews type M, 75°F working temperature, is perhaps most commonly used. It is considered good practice to maintain the room temperature for the type M wax around 75°F when recording.

The procedure for preparing the wax consists briefly of the following steps:

1. At the center of the wax, which is usually indicated by a cross mark, a 9⁄32-in. hole is drilled to a depth of ½ in.
2. A coarse cut is made for a depth of about ⅛ in. on one face of the wax and repeated as necessary to obtain a perfectly flat surface. The wax is later reversed, the first cut surface becoming the base for the finished wax.
3. On reversing the wax, a hole is cut from the other side to meet the hole drilled on the bottom.

4. A coarse cut is now made on the top surface and repeated where necessary to produce a smooth and flat surface. The wax is now ready for the final shaving or polishing cut, which is done with a sapphire or ruby cutting tool.

5. The face of the shaving knife is usually set at an angle of between 40 and 50 deg to its line of travel, depending upon the particular design of the knife. Its rounded end is toward the center of the wax. The cutting face of the knife is set at an angle of 90 deg to the surface of the wax. The turntable revolves in a counterclockwise direction.

6. The suction nozzle is placed close to the cutting knife, about ⅛ in. from the front face and $\frac{1}{32}$ in. above the cutting edge.

7. The best finishing speed is usually determined by experience, but generally ranges from 150 to 160 rpm. The finished cut on the wax should give a perfectly polished surface free from ripples or blemishes of any kind.

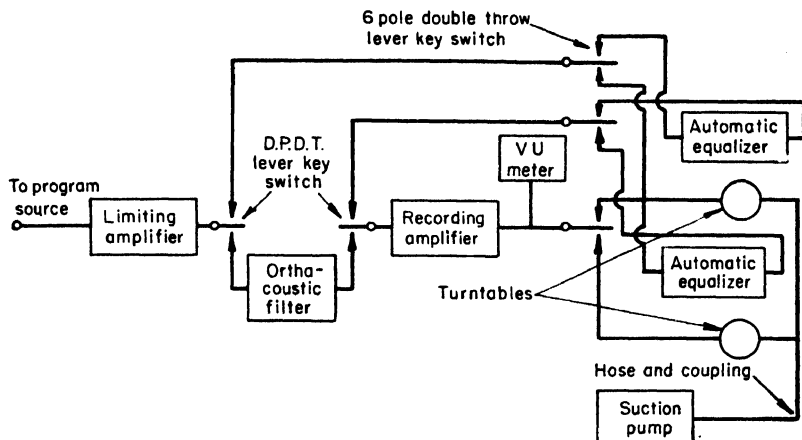



FIG. 34. Recording equipment layout for two type 73-B recorders.

31. Record Processing. Briefly, this consists of various steps, after obtaining the soft wax record, to produce the final hard record for commercial use. A complete description of each step would go beyond the limits of this section. The following are the essential steps in this process:

1. The surface of the engraved soft wax disk is rendered conductive by spreading a very thin, extremely fine conducting powder, such as metallic powder, over its surface; by the finer processes of depositing silver from a solution of silver nitrate; or by sputtering pure gold of very minute thickness on the surface. This coating is for the purpose of forming one electrode in the electroplating process.
2. Electroplating of this record with a sheet of copper $\frac{1}{32}$ to $\frac{1}{16}$ in. in thickness deposited on the wax. The negative electroplate obtained is separated from the wax and used to hot-press a molding compound, such as shellac, mixed with a finely ground filler. The first electroplate obtained is called a *master*.
3. Two test pressings are made from the first master, after which it is electroplated with a positive.
4. From this positive, sometimes referred to as an *original*, a metal mold or *stamper* record is made.
5. From the record, duplicate originals may be made and, from them, duplicate molds or stampers. By thus making a number of duplicates, it is possible to protect the original master from injury.
6. From each stamper it is possible to obtain as many as 1,000 finished pressings.

Generally, it may be said that the duplicating process reproduces everything on the original wax engraving to such a fine degree that the only difference one may observe is in the materials, one soft wax, and the other a harder, more durable plastic, composed of shellac, vinyl, or acetate compounds mixed into a filler having very little

abrasive property. The surface of these manufactured records is considerably harder than the nitrocellulose coating on metal-covered disks used for direct playback and, with a sufficiently light reproducer, will reproduce with good quality up to 1,000 playings.

32. Re-recording. It is common practice to select desired portions of a sound record by a process of re-recording. This is done with both disk and film records. Either can be played on standard reproducing equipment, which then serves as the input to the recording system, in place of the microphones. Special re-recording equipment is also used, one type consisting of a film reproducer and a film recorder combined in a single instrument and actuated by a single motor. The output of the reproducer photocell is, of course, returned to the recorder light valve in the same casing only after it has passed through an external amplifier. This instrument is used to copy on 16-mm film a sound track that was originally recorded on 35-mm stock; optical reduction, however, is also used for that purpose. Duplication of records and films is often called "dubbing."

Re-recording is used to superimpose special sound "effects" upon a record. For this purpose two or more reproducing systems are connected as a parallel input to the recorder amplifier. The method offers superior control over the relative volume of such sounds as gunshots, background music, or storms and, moreover, tends to reduce the cost of production. A library of "effect" records is maintained at many studios.

Originals intended for re-recording are sometimes made abroad by cutting a lateral track in discarded film, which is reported to be entirely serviceable for this purpose and to withstand many playbacks without damage.

33. Electrical Recording Machines. It is essential that a recording machine of a precision type should have a constant speed. For this reason it is usually driven by a synchronous motor. The mechanical inertia of the revolving table assists in keeping the rotational speed constant, the speed regulation of the disk being usually better than 0.3 per cent. It is customary to mount the driving motor on vibration dampers in such a manner as to prevent the motor vibration from reaching the revolving table. Vibration from the motor shaft is kept from reaching the shaft of the revolving table by using either belt drive, rubber differential speed rollers, or both. The spacing of the grooves cut on the disk is controlled by gear trains and the lead screw which moves the cutter head toward the center of the disk. The number of grooves engraved per inch can be set by means of the gears. A suction tube is provided for removing the shaving or thread produced while engraving. A microscope and groove illumination lamp facilitate examination of the engraved grooves. A playback pickup arm is generally provided in addition to the engraving cutter mechanism to permit playback of the record for quality checking.

34. Recording Heads or Cutters. The essential requirements of a recording head suitable for producing high-quality recordings are as follows: (1) freedom from amplitude distortion in producing undulations on the disk record, (2) suitable frequency-response characteristic over a range of 40 to 10,000 cps to produce constant-amplitude and constant-velocity recording over the frequency ranges required, (3) freedom from mechanical resonance which would tend to cause overcutting, and (4) reasonably good efficiency in transformation of complex electric wave energy into mechanical vibration of the cutting stylus.

There are numerous types and designs of cutting heads manufactured for recording sound on disk. The most common in present-day usage are the electrodynamic and the piezoelectric crystal types.

Electrical recorder heads provided for disk recording are generally designed so that the average linear velocity of the stylus (which may be expressed as a constant $\times$ the frequency $\times$ amplitude) is proportional, over a wide range of frequencies, to the impressed voltage, or $v = kfe$. The method of damping the moving system varies with different recorders.

A drawing of an electrodynamic type of recording cutter is shown in Fig. 35*a*. With a modulated current passing through the winding of this instrument, the arma-

ture produces and transfers to the cutting stylus mechanical undulations conforming with those in the electric wave, except that the amplitude is altered somewhat by mechanical and electric means. In Fig. 35*b* is illustrated a high-fidelity recording head. This cutter head utilizes a band-pass mechanical network terminated in a dry mechanical resistance material. The balanced armature is centered by means of a tempered steel spring. It is supported on knife-edge bearings upon which the lateral stylus motion is centered. Nicaloi is used for the pole pieces of the permanent magnet.

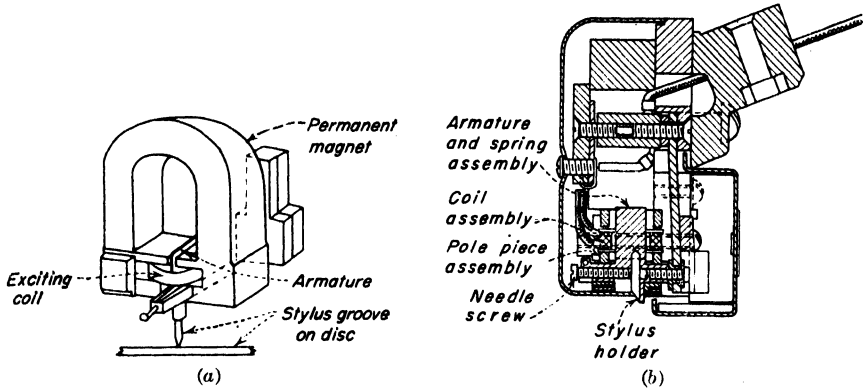



FIG. 35. (*a*) Electrodynamic type of recording reproducer; (*b*) high-fidelity recording head.

The frequency-response characteristic of this cutter head is shown in Fig. 33. Below 800 cps, frequencies are controlled to hold amplitude constant, the stylus velocity decreasing as the frequency is reduced. Above the 800-cps point the response curve shows constant-velocity motion well over a frequency of 10,000 cps. It is possible by electrical means to move the turnover point in this curve from 800 cps to a lower frequency of, say, 500 cps if desired.

While the electrical input impedance of the cutter head itself is approximately 5 ohms, an electrical-impedance compensating network can be secured to retain the total input impedance of 15 ohms throughout a wide frequency range. A high-quality amplifier having a power output of 10 watts or more is recommended for driving this cutter head.

FIG. 36. Circuit for correcting characteristic of recording head.

35. Crystal Cutting Head. This type of recording head utilizes a 5-ply Bimorph Rochelle-salt crystal to drive the sapphire stylus to engrave sound waves on disk records. For the constant-amplitude recording range the voltage applied to the crystal of the cutter head is normally 75 volts rms, while for the constant-velocity range of recording it is about 150 volts rms. Since the internal impedance of the head is rather high normally, 159,000 ohms at 100 cps, the actual power consumed by the crystal is rather small, being less than 1 watt, although the power output recommended from the driving amplifier is considerably more.

A corrective equalizer is required with the cutter for constant-velocity recording above 350 cps. Under correct operating conditions the manufacturer shows that this cutter has a frequency characteristic substantially flat within ±3 db between 30 and 10,000 cps.

A sapphire cutting stylus is recommended for use with the cutter head. For most conditions of recording the groove depth is 0.0025 in. for cutting soft wax and 0.0015 to 0.002 in. for nitrocellulose records.

36. Measurement of Frequency Response. By examination of the frequency-response curves of the various component parts of a recording system the over-all performance of the system can be checked. The program microphones and amplifiers which feed the recording head are measured in a conventional manner with a standard sound source, beat-frequency oscillator, output meter, or cathode-ray oscillograph. Under these conditions the output of the amplifier at the terminals of the cutting head is usually flat within ± 1 db between frequencies of 40 and 10,000 cps.

The recorder cutting head, however, usually has a sloping frequency characteristic (Fig. 33). The response of the cutting head alone has been measured by supplying constant level tone at various frequencies to the head and, by means of a tiny mirror attached to the stylus, reflecting a beam of light into a phototube. It is usual practice to measure the response of the cutter and disk material together.

This consists in making a recording of the output of a beat-frequency oscillator held at constant voltage at the cutter terminals. Frequencies usually recorded in order from outside to inside are as follows: 10,000, 9,000, 8,000, 7,000, 6,000, 5,000, 4,000, 3,000, 2,000, 1,500, 1,000, 800, 500, 300, 200, 150, 100, 80, and 50 cps. The completed record is then removed from the turntable; and under a concentrated single source of light, the reflection of light source as seen in the grooves shows peculiarly patterned shapes similar to their descriptive name "Christmas tree."[1] The pattern is symmetrical about the radius of the disk. It is actually a graphic representation of the frequency responses of the cutter and disk material together. The radius of the disk is the axis of frequency, the end of the pattern nearest the center being the lowest frequencies. The width of the pattern measured perpendicular to the disk radius is proportional to the undulations of the groove. This in lateral recording corresponds exactly to modulation depth. This phenomenon is due to the reflection of light over a wider band, the greater the ratio of modulated groove width to depth.

Inasmuch as good reproducing equipment usually has flat characteristics, the Christmas-tree pattern may be produced with straight sides from the turnover frequency of, say, 500 to 7,000 cps. Below this, it is customary to compensate the loss of low frequencies by boosting them with electrical filters in the reproducer. If it is noticed that pronounced peaks are in the pattern, the cutter head may be adjusted or filters inserted to produce the response characteristics required.

37. Record Reproducing Facilities. Transcribed programs generally originate in studios located separately from those in which recording is done. It is quite evident that, if full advantage is to be taken of the high-quality program material recorded on disk records, the transcription or producing equipment must also be of the precision type.

The transcription turntable is generally driven by a high-torque synchronous motor cushion-mounted within the console or cabinet. The motor shaft is flexibly coupled to the main turntable spindle. Speed regulation is reduced to a very small value for both rotational speeds of 33⅓ and 78 rpm by means of flywheel inertia and a mechanical filter on the drive shaft.

Speed reduction of the RCA type BQ2A turntable is accomplished by means of a heavy-duty ball-bearing speed-reduction mechanism operated by a button located at the rim of the turntable disk. Noise and vibration pickup is kept at a minimum by cushion-mounting the motor and spindle-housing and cushioning the suspension arms.

Special consideration is generally given to the design of a satisfactory tone arm and reproducer head for high-fidelity reproduction. The reproducer head must be light in weight and in pressure on the groove of the disk. Normally the pressure exerted by the diamond point stylus as measured by means of a spring balance or postal scale should not exceed 2 oz. A more desirable weight is less than 1½ oz. A lightweight tone arm and reproducer head assists in the reduction of record hiss or scratch noise and also the reduction of high frequencies especially near the center of the disk. Lightness also assists in securing more playbacks from a recording, since a lateral reproducer having a stylus which operates too stiffly or having too great a pressure on the disk tends to erase the higher frequencies from the record groove.

[1] See Chap. 19, Sec. 111.

Commercial reproducer heads generally utilize electrodynamic or piezoelectric principles as electric generators to convert mechanical force supplied by the groove modulation through the stylus assembly to the electrical generator element.

The reproducer of Fig. 35a is equipped with a permanent diamond point, the radius of which corresponds to the 0.0023-in. standard for lateral cut nonabrasive high-fidelity records. The armature is of the clamped-reed type. The two upper air gaps are filled with nonmagnetic material and are inactive. A linkage having a 6:1 leverage ratio is provided since the armature impedance is too high to be directly coupled to the record groove through the stylus. A diamond point is secured in the lower end of an extremely light pivot-arm spring supported vertically but rigid laterally. The pivot arm is thus permitted to rise without lifting the entire head. In the direction of useful motion transmitted to the armature the linkage has a minimum of compliance with a resultant cutoff of about 9,000 cps. This peak is reduced by means of a block of loaded rubber arranged as a selective damper approximately adjusted for the resonant frequency.

A shunt capacitance located within the tone arm is generally connected across the pickup coil to react broadly with the inductance, increasing the response through the upper frequency range. An equalizer may be placed directly at the output of the pickup head to compensate for losses in the record modulations.

The piezoelectric type of lateral disk-record pickup head utilizes a Bimorph crystal under torsional strain to convert mechanical modulations of the record groove into electrical waves. The sapphire stylus used with this reproducer is set in a small screw which fits the thread of a hollow magnesium block. The motion of the chuck is converted into a torsional strain in a bronze wire. This in turn conveys a twisting force to the Bimorph crystal sealed hermetically within a compartment. The emf produced at the electrodes of the crystal is developed from the twisting force produced by the stylus and attachment mechanism.

This type of reproducer head is normally rather light in weight, resulting in a stylus pressure of approximately 1 oz on the disk. It may be used for reproducing either constant-amplitude or constant-velocity recordings, the type of electrical compensating network required being dependent upon the particular characteristics of the recordings.

38. Orthacoustic System. There is a limitation on the amplitude of the lower frequencies recorded upon a disk. This is corrected by a sloping characteristic in the response curve below the turnover point, brought about either mechanically in the recording head or electrically with a suitable network. The undesirable needle h-f hiss is another limitation to be overcome for a satisfactory recording and reproducing system, as well as the l-f rumble caused by the turntable and building vibration, etc.

FIG. 37. Orthacoustic-recording-characteristic curve which gives preemphasis to high frequencies.

In the NBC Orthacoustic system, recording and reproducing units are individually compensated to offset characteristics of each other and thus create a reproduction which is very nearly the equivalent of the original sound. Below 100 cycles the characteristic of the recorder system is made constant-velocity by electric means. This tends to give preemphasis to the low frequencies. Then it rises from 100 to 500 cycles on a constant-amplitude basis in accordance with the mechanical and electrical characteristics of the cutter.

Above 500 cps a preemphasis above a constant velocity is given to the high frequencies, especially over the noise frequency range.

The necessary characteristic for reproduction is the inverse of this curve, or Fig. 37,

secured by electrical and mechanical means, especially those of the transcription head itself. An over-all response curve is produced which is flat over the desired range.[1]

39. Magnetic Tape Recorders.[2] The magnetic system of recording sound patterns upon a record of tape by magnetization is illustrated in a simple manner in Fig. 39.

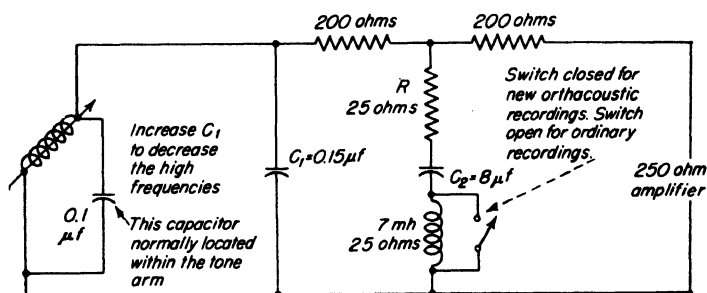



FIG. 38. Compensation filter for orthacoustic reproducer.

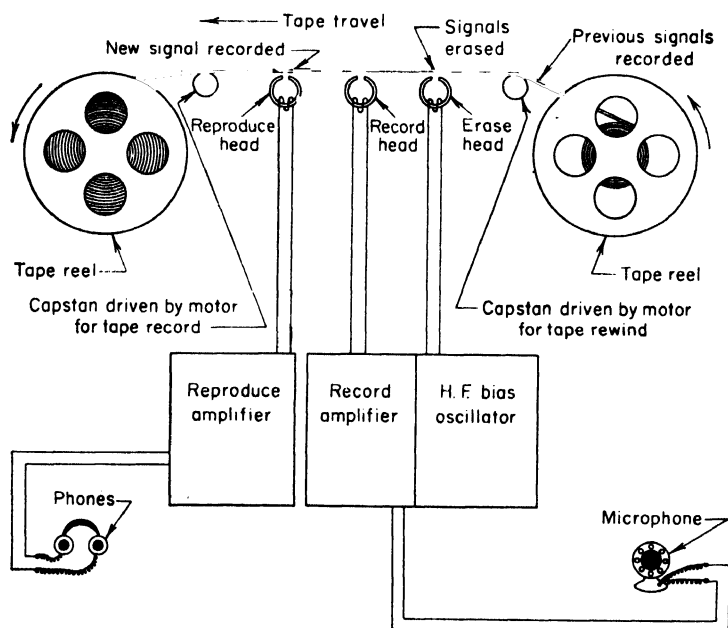



FIG. 39. Essential parts of magnetic tape recorder.

The magnetic tape moves rapidly past the pole pieces of an electromagnetic recording head while the windings are energized by the a-f signal source.[3] In traversing the windings of the recording head magnet, the audio signal currents of complex wave shape cause a fluctuating field to exist in the pole pieces. Corresponding longitudinal magnetic patterns are impressed in the moving tape.

[1] EDITOR'S NOTE: Production of the "microgroove" 33⅓-rpm records by Columbia Broadcasting System utilizing new record material virtually eliminated the noise common to the old 78-rpm disk records and, at the same time, made it possible to record upward of 20 to 30 min on a single side of a standard 12-in. disk. Extraordinary expansion of the record market for home-entertainment purposes resulted.

[2] See Chap. 19, Secs. 121 *et seq.*, of this handbook.

[3] Ranger, R. H., Magnetic Tape Recorder for Movies and Radio, *Electronics*, October, 1947.

Reproduction of the audio signal impressed by magnetism on the tape is accomplished by connecting the output of a reproducing head electromagnet through an amplifier to a loud-speaker or other electroacoustic translating device and running the magnetized tape past the pole pieces of the reproducing head in the same direction and at the same rate of speed. Under these conditions, the magnetic patterns in the passing tape create fluctuations in the magnetic field of the pole pieces and generate voltages in the windings which correspond to the wave form of the original audio signal.

Recording sound on a magnetic tape offers certain advantages over other methods especially in applications where records of a temporary nature are desired. In this case, the same medium may be used over and over again to make new recordings when old ones are no longer required. The sound record on magnetic tape may be erased and it may again be made ready for reuse by simply causing it to pass through a steady high-frequency magnetic field which causes the old signal to be erased an instant before the new signal is impressed.

FIG. 40. Tape-magnetization process.

Another advantage is in making continuous long-duration recordings upon a single magnetic tape medium, thus avoiding the procedure of having to change recording disks at intervals when using the sound-on-disk system. A sound record on magnetic tape may be stored away for an indefinite time without fear of deterioration. The tape record may be played back hundreds of times without appreciable loss of quality, which is an advantage over disk recordings which depreciate rapidly when used continuously.

40. Magnetic-recording Medium. The recording tape used for magnetic recording consists of finely divided iron oxide particles deposited on a plastic strip having dimensions usually about $\frac{1}{4}$ in. wide and 0.002 in. thick. The strip is manufactured of a plastic compound which has excellent properties of strength and durability to withstand the mechanical stresses produced during the recording and reproducing processes. For the purpose of illustration, assume that the tape is moved through a field in which a magnetic force alternates in accordance with a sine wave as illustrated in Fig. 40. The molecules of the iron oxide particles are oriented according to the direction and magnitude of the magnetic fields. The pattern of the audio signal stored in magnetic form on the tape is shown.

FIG. 41. Magnetization characteristics of magnetic tape.

The magnetization curve of the iron oxide particles of the tape is similar to that in Fig. 41. It should be noted that near the origin of the curves a very nonlinear condition exists. The signal recorded upon the tape under these conditions would not be directly proportional to the signal voltage applied to the windings of the recording head and a high amount of distortion would result when the signal was reproduced.

By superimposing a high-frequency constant-amplitude bias signal upon the audio signal to be recorded, this distortion can be reduced considerably. Usually the frequency of this bias voltage is selected to be about five to seven times the highest

a-f limit of the recorder. This is necessary to prevent beating between the bias and the harmonics of the recorded signal. For example, a magnetic recorder capable of reproducing up to 15 kc would require a bias voltage having a frequency of about 100 kc. During the instant of time that the tape is in the gap of the recording head, the bias voltage causes the magnetizing characteristics of the iron oxide to follow the dashed-line loops, as illustrated in Fig. 41. These are sometimes referred to as *minor hysteresis loops*. Coinciding with the movement of the tape from the gap, the effect of the magnetic field caused by the h-f bias diminishes to zero. The iron oxide coating of the tape then assumes a permanent magnetization. This is known as *remanent induction*, which is determined by the gap flux at the instant this particular part of the tape leaves the gap. Following the recording process, the tape contains a pattern of magnetic flux proportional in magnitude and direction to the audio signal recorded.

41. Magnetic Recording Head. The magnetic field is produced in the gap of the recording head over which the tape passes. An incomplete ring-shaped electromagnet forms the recording head, which is usually of the appearance shown in Fig. 42. The gap is formed by the discontinuity of the ring, which is made of high-permeability material to form the core of the electromagnet. The gap is usually filled with a nonmagnetic substance to form a smooth surface over which the tape passes.

FIG. 42. Recording head.

42. Magnetic Reproducing Head. The electromagnetic head used for reproducing the audio signal is quite similar in construction to the recording head. When the tape is moved past the gap of the reproducing head, the magnetic flux on the moving tape induces a voltage into the winding of the reproducing head. This is proportional to the number of turns on the winding and the rate of change of flux as produced by the moving tape.

For most satisfactory results, the gap in the reproducing head should be as small as possible. A good rule is to make it of a physical length equivalent to less than a wavelength of the highest frequency to be reproduced, which would be less than 1 mil. As the gap is decreased in width, however, the induced voltage is decreased and for this reason there is a practical limit to the minimum width of gap to maintain a signal-to-noise ratio that is of adequate proportions.

The audio signal induced across the windings of the reproducing head may be calculated from the following:

$$E = B_m V \sin (\pi\omega/\lambda)$$

where E = voltage induced
B_m = maximum flux density of recording material
V = velocity at which tape moves across reproducing head
ω = width of head gap
λ = wavelength of the signal on the tape

This expression indicates that the voltage produced across the windings of the reproducing head will increase directly with an increase in velocity of the tape and with a decrease in the wavelength.

When the velocity of the tape and the gap width are constant, the voltage delivered from the windings of the reproducing head will be directly proportional to the frequency under conditions where the wavelength on the tape is large compared with the width of the gap, and as a result the curve of output voltage vs. frequency is shown by the solid line A in Fig. 43. Because of an increase in the electrical losses of the core material as the wavelength nears the dimensions of the reproducing head gap, the actual output voltage drops off at the higher frequencies and follows the curve shown by the dashed line.

The problem of providing an over-all flat frequency-response characteristic for the

recorder is solved by an equalization circuit.[1] Series R and C of proper values connected across the input of one of the earlier stages of the reproducing amplifier as a solution provide an h-f droop characteristic that is inverse to the characteristic of the reproducing head.

The h-f response characteristic of the recorder is extended further by providing additional equalization in the recording amplifier. This is in the form of a frequency-boost circuit designed to compensate the dropping off of the high-frequency end of the reproducing-head characteristic because of excessive core losses.

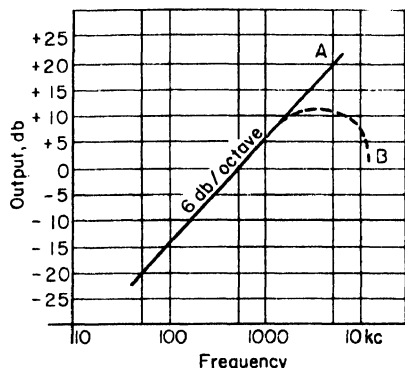



Fig. 43. Output characteristics of the reproducing head.

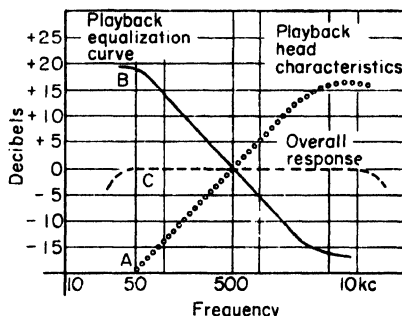



Fig. 44. Equalization and over-all response of reproducing system.

43. Magnetic-recorder Components. The parts that are essential to a tape-recording setup are: (1) the tape and its transport mechanism including the driving motor; (2) the erase, record, and reproduce heads; (3) the h-f bias oscillator; (4) the record amplifier; (5) an audio input source; (6) a loud-speaker or telephones for monitoring; (7) a cabinet; (8) a power supply; and (9) miscellaneous mechanical controls and electric switching systems.

Magnetic tape is available upon standard-size plastic reels having nominal diameters of 7 and 10½ in. which contain the tape in lengths of 1,200 and 2,400 ft, respectively. The standard tape speeds are 3¾, 7½, and 15 in. per sec. The speed of 3¾ in. per sec will provide in high-quality recorder units a satisfactory response up to 7,500 cps. This slower speed has the advantages of a longer "full-track" playing time of 2 hr and 8 min for a tape length of about 2,400 ft. It is available on a 10½-in.-diameter standard reel. At a tape speed of 7½ in. per sec it is possible to secure acceptable response up to 10,000 cps with a full-track playing time of 1 hr and 4 min per 2,400 ft of tape and for a speed of 15 ft per sec good response of 15,000 cps is obtainable with a full-track playing time of 32 min.

The tape is used full-track where one signal only is recorded on the full length of tape. The tape can also be used "half-track" by recording the signals on both sides simply by reversing the tape on the reels. If half-track operation is used, twice the playing time becomes available for a given length of tape but there is, of course, the disadvantage of having to reverse the tape on the reel to play the entire length of the recorded signal.

The tape is moved by a capstan driven by a mechanical speed reduction, usually with belts and pulleys, from the shaft of a synchronous motor. It is customary for the motor to be of multiple speeds such as 1,800 or 900 rpm, thus simplifying the problem of changing the tape speed. A braking system is required on the tape-transport system to maintain a constant tension in the moving tape. The over-all speed regulation of the transport system is required to be a small fraction of 1 per cent to provide satisfactory operation of the recorder unit with respect to "flutter" and

[1] Korner, Otto, Frequency Response of Magnetic Recording, *Electronics*, August, 1947.

"wow" defects. These are usually in the order of 0.3 per cent rms or less. The flutter and wow measurements include all components between the frequency range of zero to 300 cps using an rms meter calibrated to read the peak value of a constant-amplitude sine-wave flutter. In the process of rewinding the tape on a reel, facilities are provided to increase the speed tremendously to effect a saving in time.

One of the keys to the design of a satisfactory tape-recorder unit lies in the recording and reproducing heads. They must be constructed in such a way as to record and reproduce the magnetic patterns on a tape with minimum distortion and maximum signal-to-noise ratio. For full-track operation, the peak record level is defined as that in which the over-all (input to output) rms harmonic distortion does not exceed 3 per cent when measured on a 400-cps tone assuming that noise is measured when erasing a signal of peak recording level in the absence of new signal. Thus, the bias and erase noise are included, as well as playback amplifier noise. Furthermore, it is assumed that all frequencies between 50 and 15,000 cps are measured.

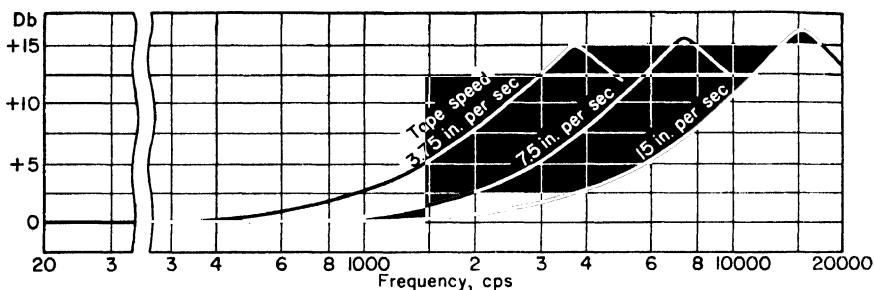



FIG. 45. Relation between tape speed and frequency response.

For full-track operation, the peak record level as compared with unweighted noise should be at least 50 db for a tape speed of 3¾ in. per sec to 60 db for a tape speed of 15 in. per sec.

The reproducer head is a most critical part of the tape-recording unit because signals recorded on tape are of low intensity. It requires approximately 60 db of amplification to bring the peak recorded signals up to a level of zero dbm. Inasmuch as the dynamic range possible with tape recorders approaches 60 db, the weakest signals are nearly 130 db down or on levels that are slightly higher than the thermal noise in some vacuum tubes.

While the record head is generally constructed quite similarly to the reproducer head, a key to proper design lies primarily in the width of the gap in the high-permeability ring. This problem has been explained in Sec. 41.

The erase head contributes to satisfactory operation of a tape recorder by providing the removal, as nearly as possible, of the magnetic patterns previously recorded. The h-f field supplied by the bias oscillator provides sufficient energy to cause the erase head to remove by magnetic saturation the audio signals to be erased.

The record amplifier of a high-quality tape-recording unit must have high gain, a high amplified signal-to-noise ratio, and a low inherent amplitude distortion. Equalizers giving the required over-all frequency response with preemphasis on the higher frequencies are required to compensate h-f losses caused by magnetic saturation of the iron oxide particles as the tape passes through the recording-head gap.

A typical characteristic of a record amplifier is shown in Fig. 45. It should be understood that the separate equalizers are switched into the circuit to conform with the tape-speed requirements.

A bias signal of a frequency in the order of 70 to 100 kc is combined with the audio signal being recorded to improve the magnetic remanence of the magnetic coating on the tape. The ratio of the intensity of the bias signal to the audio signal is usually 1:2 or less.

Excellent characteristics of the reproducer amplifier are required to provide high gain, high signal-to-noise ratio, and low inherent amplitude distortion to the amplified audio signal that passes through it. The amplifier must have an equalizer in one of the earlier stages to give preemphasis to the lower frequencies to offset the natural characteristic of increasing voltage output with respect to frequency caused by the tape passing through the reproducer head. The over-all effect required to provide a flat over-all frequency-response characteristic for the recording unit has been explained previously and is illustrated by the curves shown in Fig. 44. The typical frequency-response characteristics of the reproducer amplifier of a high-quality tape recorder are shown in Fig. 46.

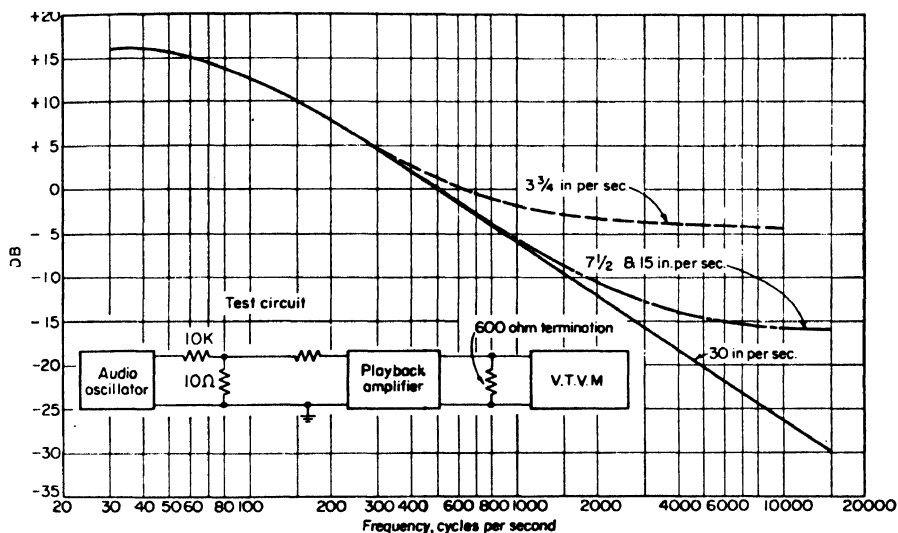



FIG. 46. Playback amplifier response magnetic tape recorder.

44. The Complete Magnetic Tape Recorder. In Fig. 47 is shown a schematic diagram of the electronic circuits which comprise the Ampex Model 350 magnetic tape recorder. This unit will accommodate tape speeds of 3¾, 7½, and 15 in. per sec. The over-all frequency-response characteristics are (1) ±2 db over the frequency range of 50 to 7,500 cps at tape speed of 3¾ in. per sec; (2) ±2 db from 40 to 10,000 cps for a tape speed of 7½ in. per sec; and (3) ±2 db from 30 to 15,000 cps. The peak record signal-to-noise ratio at full-track tape speeds of 7½ and 15 in. per sec is 60 db and for half-track it is 55 db. The other important operational characteristics are within the ranges previously explained.

45. Wire Lines. Wire telephone systems are employed almost exclusively for the national distribution[1] of programs to the various stations connected on a network.

The frequency band which is transmitted over long-distance program circuits extends from about 100 to about 5,000 cps; to transmit music with improved fidelity a wider band than the above is desirable. A few circuits are at present available which extend the band down to 30 or 50 cps and extend the higher range by 2,000 or 3,000 cps. Program transmission circuits must be designed to handle wide ranges of volume. At present the volume range is limited to some 25 or 30 db, from about +8 vu down to about −22 vu. Obviously, since the dynamic range of a symphony orchestra is about 60 db, the wire-line circuit necessitates some compression of the dynamic range, especially on long network circuits.

46. Standardization of Transmitting Levels. To obtain optimum conditions from the standpoint of noise and cross talk, it is desirable to transmit program material into

[1] Clark, A. B., Wire Line Systems for National Broadcasting, *Proc. IRE*, **17**, 1998, November, 1929.

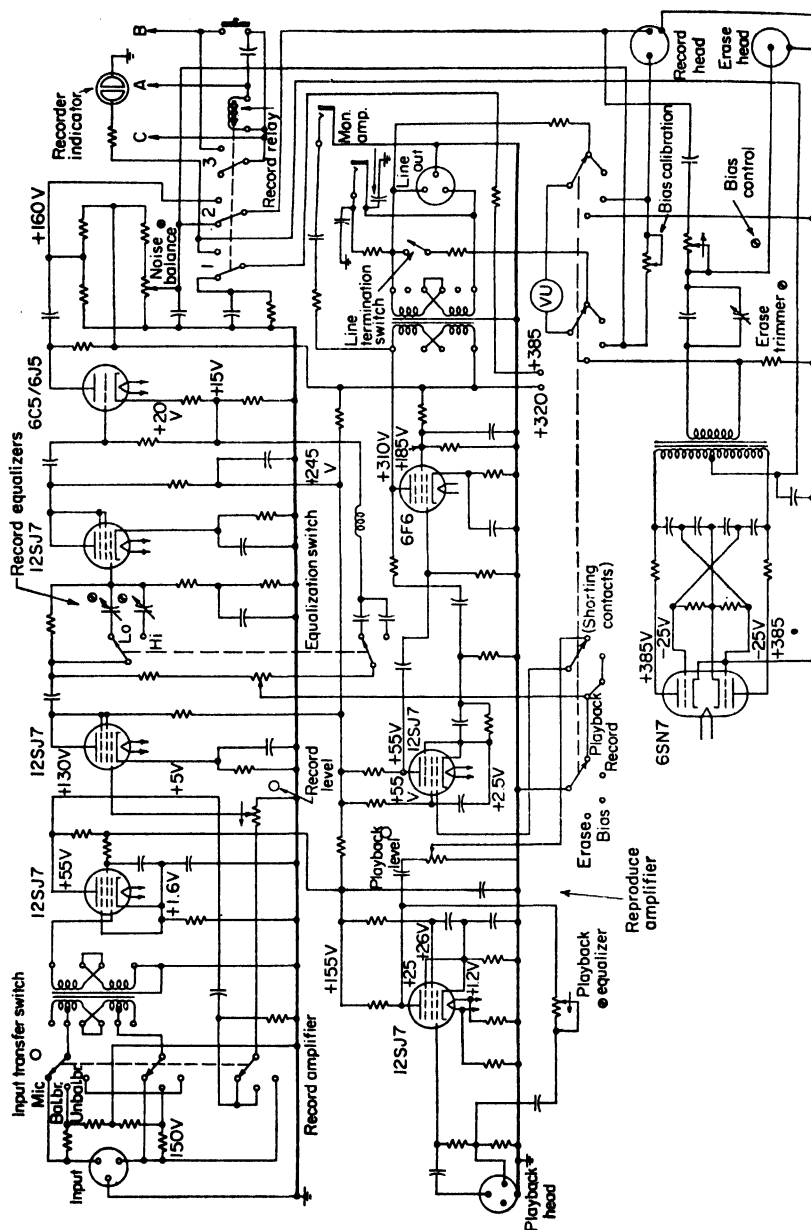



FIG. 47. Simplified schematic diagram of Ampex Model 350 magnetic tape recorder.

loops at as high volumes as practicable. Telephone-company experience has demonstrated that in general +8 vu is about the highest volume of program material that can be tolerated in a local cable plant of the kind in which broadcasting loops are routed, from the standpoint of interference with other circuits. In view of these conditions, therefore, +8 vu (+14 vu output of amplifier followed by a 6-db pad) is

the standard volume level for transmitting to loops in local telephone cables. The isolating pad is for the purpose of isolating the amplifier from the telephone-company loops.

RADIO FACILITIES

47. A-f Equipment. The process of transferring programs from the main control room of the studios to the broadcast transmitting station is generally accompanied by a considerable reduction in the program signal level. Attenuation caused by the wire line upon which is added that caused by the line equalizer lowers the signal intensity as much as 25 db. A line equalizer consists of a specially designed network containing correctly proportioned values of L, R, and C. Irregularities in the wire-line frequency characteristics are smoothed out by the equalizer to produce a uniform frequency response of the wire line over as wide a range as practicable.

To increase the level of the incoming signals to a sufficient intensity to drive the first tube of the speech amplifier of a broadcasting transmitter, a line amplifier is required. This amplifier is usually of a high-quality limiting type having sufficient gain to raise the audio program signal to a level of approximately +15 vu. At this level it enters the first speech-amplifier stage. The line equalizers, line amplifiers, variable attenuators, volume indicators, monitoring amplifiers, microphone for making local announcements, together with their switching equipment and jack panels, are normally mounted on shielded racks in a room called the *control room*. The shielding of the room itself sometimes consists of an outside-grounded copper screen containing within it a floating copper screen.

48. Limiting Amplifier. A special type of amplifier normally used in the speech-input layout at the broadcasting transmitter is of the compressing or limiting type. This amplifier automatically reduces the channel gain whenever the program peaks become excessively high. Thus it tends to prevent overmodulation. As a result, distortion due to transmitter overmodulation can be avoided while at the same time the average modulation can be raised with a corresponding audio power gain at the receiver. This is noticeable especially at low passages of program material where background noise may become objectionable.

By rectifying a small portion of the program signal output, a bias voltage control is provided on a program signal amplifier. This action does not just cut off the program peaks, but it reduces the gain and then allows it to rise slowly to normal.

The signal voltage is amplified and then rectified in a diode with a result that a variable d-c bias voltage appears across a resistor in series with the bias voltage to the grids of the first stage of the amplifier. With an increasing signal, the bias becomes more negative and the output of the amplifier is reduced. This action does not occur, however, until the audio signal level applied to the control tube exceeds the fixed bias of this tube.

A potentiometer across the secondary of the input transformer is utilized as a variable-input control from which the corresponding input level at which the compression takes effect is varied. Owing to the high gain of the amplifier (54 db), the beginning of the compression may be as low as −40 vu. Provision is also supplied for adjustment of the output of the amplifier by means of a potentiometer in the input of the second amplifier stage. By means of this control the output level can be set anywhere within the range of −30 to +15 vu.

To compress sudden peaks of the program wave, the control circuit must function very quickly. The time constant of the circuits involved is such that the reduction in gain occurs in 0.001 sec. To prevent the gain from fluctuating at low audio or syllabic frequencies, a delay circuit is provided to allow the compression bias voltage applied to the grids of the tubes in the first stage to leak off slowly and return the amplifier gain to normal in about 7 sec. This delay has been set by listening tests to prevent introduction of distortion or to destroy speech inflections.

The amplifier has an output of +30 vu with 18 vu compression. The frequency response is flat within ±1 db from 30 to 10,000 cps.

49. Program monitoring facilities are a very essential part of broadcast station equipment. In broadcasting technique, *program monitor audio* refers to a monitoring

check on the audio signal input to the transmitter, whereas *program monitor radio* refers to a check on the demodulated signal secured by rectification of the carrier envelope as produced at the broadcast transmitter output. By switching from the *input signal* to that produced by rectification of the modulated transmitter carrier, the station personnel can determine by listening tests and measurements the relative

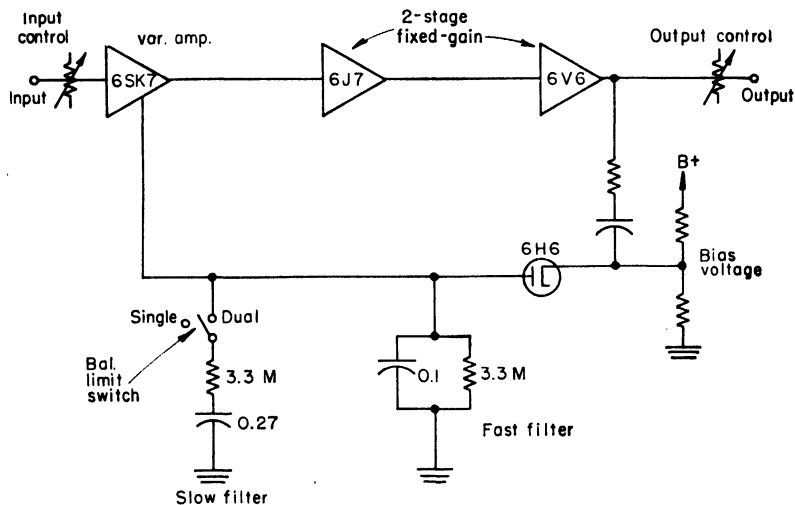



Fig. 48. Limiting amplifier, RCA BA-6A.

amount of distortion produced in the broadcasting station equipment. For monitoring the outgoing program the personnel normally listens to the program monitor radio as produced by demodulation of the signal at the antenna system. This ensures that all portions of the audio and radio transmitting equipment, as well as the antenna system, are functioning. This is indicated by monitoring loud-speakers or oscillographs.

Facilities for program monitoring are provided in a room suitably constructed and acoustically treated to provide a favorable place for listening tests in the judgment of quality. This may be either the transmitter room itself or an adjoining room called the control room where the speech input is normally located. The equipment for monitoring the audio signal consists of high-quality audio amplifiers, the gain of which can be regulated for proper signal volume; high-quality loud-speakers; and associated switching equipment. The frequency response of the entire system should be flat over a range of between 30 and 12,000 cps and higher. Additional to this equipment for program monitoring radio is a well-designed monitoring rectifier capable of demodulating the carrier signal as picked up at either the output tank circuit of the radio transmitter or at the antenna, preferably the latter.

Fig. 49. Antenna monitoring rectifier circuit.

Schematic diagrams of two types of antenna monitoring rectifiers, shown in Figs. 49 and 50, illustrate single-ended and push-pull types, respectively. These rectifiers are equipped with circuits enabling them to be used as the remote antenna current-meter rectifier, to close a carrier-on relay or time-outage clock relay as well as the monitoring signal for oscillograph or loud-speaker. In coupling such rectifiers as shown to an antenna circuit, precautions are usually taken to prevent the generation of even and odd r-f harmonics into the antenna circuit as produced by rectification. Under certain conditions such harmonic generation and radiation from the antenna system may create interference on the harmonic frequencies. For this reason, the push-pull type when inductively coupled to a high current point of the antenna system

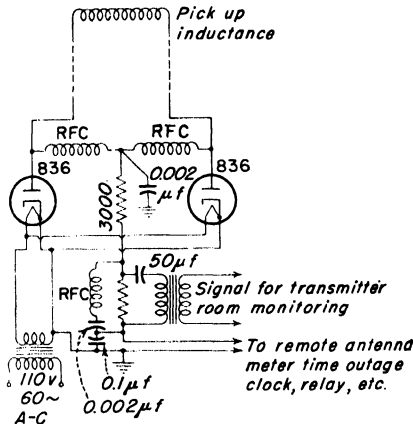



FIG. 50. Push-pull antenna monitoring rectifier.

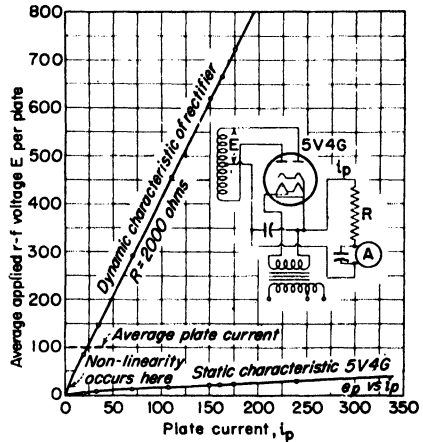



FIG. 51. Characteristic of monitoring rectifier; i_p is total plate current.

has considerable advantage over single-ended types, in that even harmonics are not as pronounced.

For rectifying the envelope of a carrier wave to secure a signal for loud-speaker monitoring or for modulation measurements with an oscillograph, it is essential that the linearity characteristics of the monitoring rectifier between the impressed voltage and the plate current be substantially straight throughout the operating range. The unit must also have a uniform frequency-response characteristic to provide reproduction of the signal without frequency distortion.

Diode rectifier tubes are used extensively for monitoring radio telephone signals. As an individual element of the monitoring rectifier, the diode itself is not a linear device since the internal resistance of the diode decreases as the anode voltage is increased. The selection of diode tubes having low internal voltage drop and the introduction of sufficient resistance in the plate circuit are required in the design of a monitoring rectifier of satisfactory linear characteristics. Linearity may be further improved by application of a constant positive bias in the plate circuit so that the diode draws steady plate current over the most nonlinear lower portions of the curves. In Fig. 51 these design features are illustrated for a 5V4G diode, which is a particularly good type for monitoring rectifier use owing to its low internal voltage drop. Tubes of higher inverse peak voltage are often required for rectifiers of higher power-handling characteristics and to withstand voltage surges (such as those caused by lightning) from an antenna circuit.

The percentage distortion of a rectifier may be approximately calculated from the dynamic characteristic by using a formula similar to that used in calculating percentage distortion of triode amplifier tubes.

50. Frequency Monitor. This instrument is required to measure the amount of deviation of the carrier from the assigned frequency. The FCC rules under Sec. 3.59 state that the operating frequency of each a-m broadcast station shall be maintained within 20 cps of its assigned value. Under Sec. 3.60 of the FCC rules, the frequency monitor is subject to FCC approval with respect to stability and accuracy. The General Radio type 1181A frequency monitor illustrated in Fig. 52 is an approved type operating within the specified limits. It contains a frequency-standard oscillator utilizing a quartz crystal together with a tube (6SJ7) in a circuit having excellent frequency stability.

The oscillator drives a 6AC7 buffer amplifier with very light coupling between them. The output of the buffer is coupled to the grid of a 6SQ7 mixer, where it is mixed with voltage from the broadcast transmitter. This voltage is picked up at some stage of the transmitter below the one modulated, is amplified by a 6AC7, and coupled capacitively to the mixer. The oscillator is adjusted to a frequency of 1,000 cps off that of the transmitter carrier. Hence the output of the mixer is a difference frequency of 1,000 cps plus or minus the transmitter deviation frequency. The frequency thus produced is amplified and passed through a 6SQ7 clipper-limiter tube to produce a square wave of constant peak-to-peak amplitude.

The square-wave voltage is amplified by a 6V6 tube circuit and passed to a voltmeter. This latter instrument is a full-wave differential diode voltmeter which indicates the potential difference between the mid-points of two series resonance circuits. In this manner equal potentials are indicated at exactly 1,000 cps, and currents flow at other difference frequencies actuating the meter accordingly over the normal range of ± 30 cps deviation which is the direct scale reading of the meter.

51. Modulation Monitor. Section 3.56 of the FCC rules requires that each broadcasting station shall have an approved modulation monitor in operation at the transmitter to measure the degree of modulation of the transmitter and for furnishing instant warning when the degree of modulation exceeds a selected specified value.

In Fig. 53 illustrating circuits of a modulation monitor, the a-m r-f signal to be monitored impressed across diode 6AL5 is thus rectified. The average value of d.c. is indicated on the carrier meter. This is proportional to the average carrier voltage. The average component of voltage across a load resistor in the 6AL5 excites two indicating devices: (1) the meter M_2 calibrated to read modulation percentage and decibels directly and (2) a flasher and alarm circuit for providing a warning when the degree of modulation is exceeded.

The modulation-indicating meter is excited in the following manner: The audio component secured from the first diode is rectified by the second 6AL5 diode detector and charges C_2. The voltage across C_2 is impressed across the first grid circuit of the 12AU7 voltmeter stage which has the indicating meter M_2 in its plate circuit. Circuit constants are made such in this instrument that the a-f peaks on the r-f carrier are as indicated by meter M_2. The neon flasher is operated by the relay tube, an 884, which is in turn driven from the first tube under the same audio component from the tube that is used for operating the indicator meter system. If desired, the instrument may be used to operate an auxiliary alarm when the modulation peaks rise to an excessive value.

Modulation indicators are usually calibrated by means of a pure sine-wave modulating signal applied to an accurate cathode-ray oscillograph and checked against the indicator. The frequency response must necessarily be flat over the audio range used to ensure accuracy of measurement over the range.

According to Sec. 3.55 of the FCC rules, a licensee of a broadcast station will not be authorized to operate a transmitter unless it is capable of delivering satisfactorily the authorized power with a modulation of at least 85 per cent. When the transmitter is operated with 85 per cent modulation, not over 10 per cent combined a-f harmonics shall be generated. Under Sec. 3.46 (FCC rules) design recommendations state that total a-f distortion from microphone terminals, including microphone amplifier, to antenna output should not exceed 5 per cent harmonics (voltage measurements) when modulating from 0 to 84 per cent and not over 7.5 per cent harmonics (voltage measurements of arithmetic sum) when modulating 85 to 95 per cent (distortion shall be

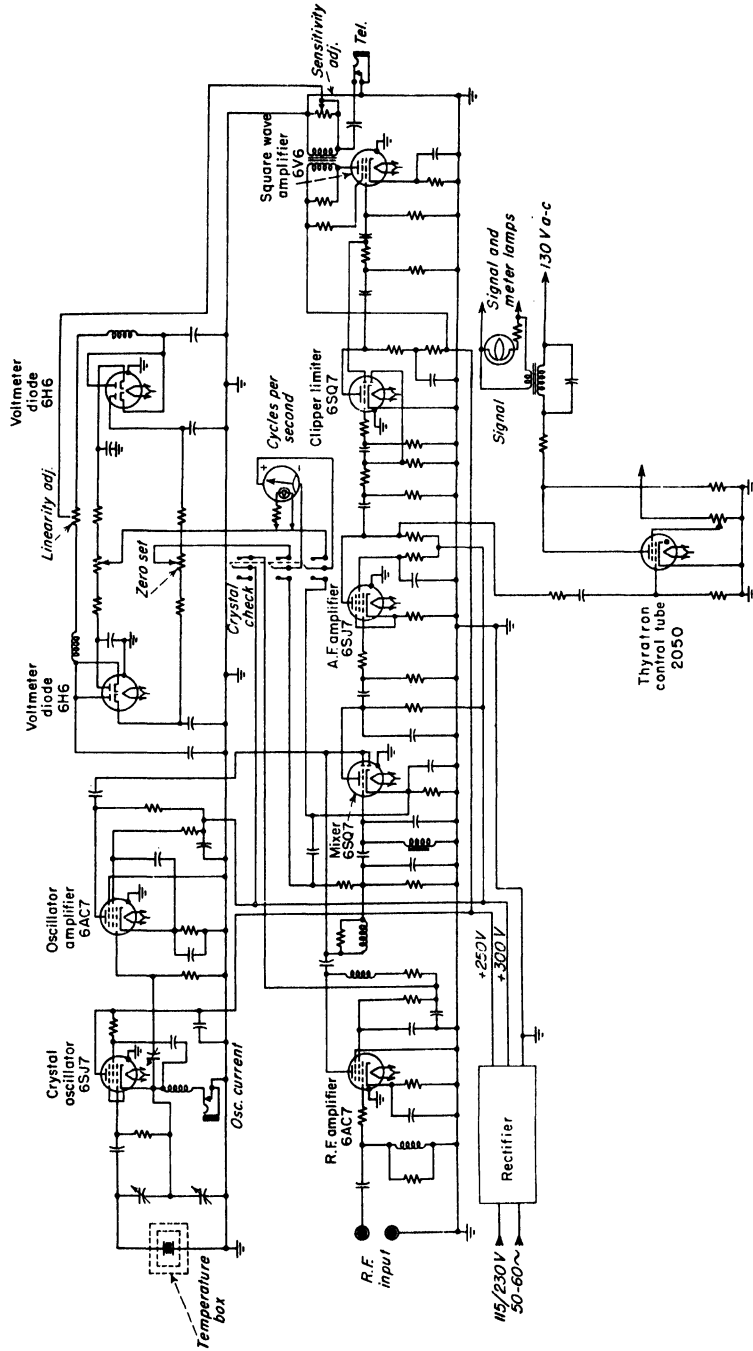



FIG. 52. Schematic General Radio type 1181A frequency-deviation monitor.

FIG. 53. Simplified schematic diagram of RCA BW66E a-m monitor.

measured with modulating frequencies of 50, 100, 400, 1,000, 5,000, and 7,500 cps up to the tenth harmonic or 16,000 cps or any intermediate frequency that readings on these frequencies indicate is desirable).

The operating percentage of modulation of all stations is normally maintained as high as is possibly consistent with good quality transmission and good broadcast practice.

RADIOBROADCASTING TRANSMITTERS

Production of a broadcasting signal that will afford a means for conveying speech and music to a broadcast listener involves the generation of a constant r-f carrier upon which there are superimposed audio frequencies in the form of side bands, the intensities of which conform as nearly as possible with those contained in the sound produced in the studio. The production of such a signal may be accomplished by several methods[1] of modulation.

In American broadcasting technique the amplitude system of modulation is used exclusively in the band 550 to 1,600 kc. The advantage of a.m. for transmission in this band lies in the production of a modulation envelope containing but a single pair of side bands, thus permitting station channel separation of 10 kc.

Compared with the a-m system, phase and frequency modulation produce an infinite number of side bands. It is evident that greater channel separation is needed and for this reason f-m stations have been assigned to the v-h-f part of the spectrum.

The primary requisites of a transmitter satisfactory for the band 550 to 1,600 kc under the present rules of the FCC are as follows:

1. Satisfactory carrier frequency stability well within the allowable FCC tolerance of ± 20 cps maximum deviation.
2. Amplitude and frequency response characteristics providing low over-all signal distortion.
3. Suitable safety devices to avoid hazards to operating personnel and electrical circuits and equipment complying with the National Electric Code.
4. Minimum carrier noise level, approved electrical metering facilities; minimum r-f harmonic frequency power output; and freedom from parasitic frequency emissions.
5. Low operating costs requiring an over-all high-operating efficiency with respect to power input, low approved power-tube operating expenses, and economical operating personnel requirements.
6. Durability, simplification of adjustment, and maintenance (requiring accessibility for repairs).
7. Reliability of service providing for continuous operation with a minimum of interruptions at rated carrier power output, modulated within legal limits.
8. Satisfactory dimensions for given power output providing for minimum installation and building costs.
9. Low initial transmitter and installation costs.
10. A pleasing appearance.

A trend is toward transmitters having high-level modulated and high-efficiency linear power amplifiers for producing the desired high-quality broadcasting signal with minimum operating expense.

52. International Broadcasting. Transmitters for this service are operated at high frequencies and for this reason are considerably different in design from transmitters operated in the 550- to 1,600-kc band. They are used with directive antennas having a power gain of 10 or more, and have carrier powers up to 200 kw 100 per cent modulated.

R-F CIRCUITS

53. R-f Amplifier Neutralization. One of the essential adjustments in an r-f amplifier circuit to obtain stability and prevent self-oscillation is accomplished through neutralizing the grid-to-plate capacitance of the triode power tubes.

For the purpose of neutralizing an amplifier stage such as the final r-f power amplifier first remove plate voltage from it and apply normal r-f excitation to the grid circuit. Tune

[1] Roder, Hans, Amplitude, Phase and Frequency Modulation, *Proc. IRE*, **19**, 2145, December, 1931.

the grid circuit to resonance in the usual manner. Next connect a low-power (5- to 10-watt) high-resistance lamp across one or two turns of the plate tank inductance. The leads to the lamp should be very short and provided with clips for convenience. Next tune the plate tank circuit to resonance with the grid exciting voltage frequency as indicated by maximum brilliance of the lamp. The circulating current in the plate tank circuit which lights this lamp includes the coupling effect of the grid-plate capacitance of the tube.

The neutralizing voltage of opposite polarity is obtained by connecting to the opposite end of the grid or plate tank circuits, as the case may be. The magnitude of the voltage used to neutralize the grid-plate capacitance current is regulated now by adjusting a neutralizing capacitor. As the neutralizing capacitor is varied, the lamp will change brilliancy, and, when correct balance is obtained, the lamp will be at practically zero brilliancy. As neutralizing capacitance is changed, some slight corrections in plate-tank tuning and in grid tuning may be necessary, owing to interactions of these two circuits. Always tune to resonance by maximum lamp brilliancy and neutralize for minimum brilliancy.

When best results are obtained by the lamp method, remove it from the plate coil, and, if more accurate adjustment is required, a low-range r-f ammeter should be inserted in series with the tank circuit. By using a meter, maximum accuracy is obtained by tuning the circuit to obtain absolute minimum current.

Since the effect of coupling between successive stages greatly affects the neutralizing, the adjustment should be made with all circuit conditions and couplings as nearly final as possible.

An entirely different neutralizing system consists of an effective inductance shunting the grid-to-plate capacitance of the amplifier tubes. Suitable d-c blocking capacitors are provided to prevent the plate voltage from reaching the grid through this neutralizing inductance.

Neutralizing adjustments with this shunt inductance may be accomplished with a high-resistance lamp or thermomilliammeter attached to the output tank circuit in much the same manner as was described for capacitor neutralization except that neutralization is accomplished by adjustment of the shunt inductance. This system has great advantages over the neutralizing capacitor method especially where it is desirable to keep circuit tank capacitance and the corresponding kva/kw ratio to a low value. This is the case where stabilized degenerative feedback is applied through an amplifier stage wherein a minimum phase rotation with frequency is required.

FIG. 54. Theoretical curves showing push-pull class B amplifier operation.

54. Class B Linear R-f Amplifiers. The operation of a push-pull class B r-f amplifier may be understood by a study of Fig. 54. Here it is shown that plate current drawn by the tubes is very closely a linear function of the grid-voltage swing. The associated output-circuit loading is adjusted so as to realize from the tube a maximum conversion efficiency. Some curves showing how plate-current efficiency varies with effective load impedance are shown in Fig. 55. The crest position on these curves depends upon the tube characteristics and the power factor of the circuit into which it operates. These curves were taken at a broadcast frequency by varying the load

upon the output circuit of a linear amplifier stage and measuring the efficiency of the stage at various d-c plate voltages.

Where the conversion efficiency is a linear function of the grid swing, the power output is necessarily proportional to the square of the grid swing. Hence the peak power output at 100 per cent modulation is four times that at which the modulation is zero. The steady power output under conditions of sustained 100 per cent modulation is 1.5 times the output of zero modulation. Therefore, in considering power-tube requirements for a class B linear stage, provision must be made with respect to filament emission and plate dissipation so that the tubes are capable of supplying peak power outputs of four times that of the nominal carrier-power output rating of the transmitter. This assumes that the modulation capability of the transmitter is 100 per cent.

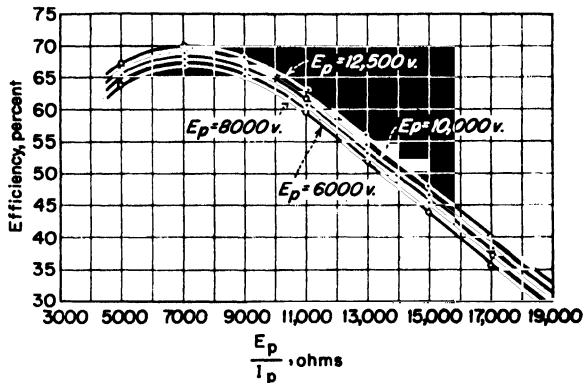



FIG. 55. Load characteristics of push-pull 892 power-amplifier tubes.

In adjusting a push-pull linear amplifier, both sides of the circuit must necessarily have very nearly identical operating conditions with respect to grid swing and circuit adjustment, so that equal plate currents are measured on the individual tubes identified as *A* and *B* in Fig. 54. The grid-bias adjustment necessarily depends directly upon the plate voltage used since the position of the characteristic curve is moved with each corresponding change in plate voltage.

As illustrated in Fig. 54, with a simple triangular wave form, the method of determining optimum grid bias depends upon the point where an extension of the straight portion of the curve intersects the horizontal axis. Both dynamic curves of tubes *A* and *B* have their straight portions in direct alignment. Distortion due to the lower bend in each characteristic curve is averaged out together with the kva/kw inertia effect in the output-tank circuit. On the other hand for maximum modulation peaks with output increasing as the excitation voltage is increased, there is a limit to the output as represented by the upper bends, points S_1 and S_2, on the curves where the tube saturation points begin.

Linearity is therefore dependent upon grid bias, grid-exciting voltage, and output-tank loading. The procedure for setting taps for correct output-tank loading consists of first saturating the grids of the amplifier tubes with sufficient r-f grid driving power. Then with one-half normal class B r-f operating d-c plate voltage applied, the amplifier is loaded until it delivers rated carrier power normally to the antenna, the plate efficiency being usually between 65 and 70 per cent. Then the grid-exciting voltage is reduced (usually by means of grid-loading resistors) until the amplifier stage with full d-c plate voltage applied delivers the same rated carrier output with a corresponding plate efficiency of very nearly 35 per cent. This is the plate efficiency for a class B r-f amplifier as specified by Sec. 3.52 of the FCC rules in the determination of carrier out-

put power by the indirect method. Under this same section the plate efficiency for plate-modulated class C r-f operation of the last radio stage as measured by the indirect method is 70 per cent for transmitters having carrier power output up to 1 kw and 80 per cent for 5 kw and over.

55. High-efficiency linear-power amplifiers are a result of work to reduce the expense for operating power of broadcasting transmitters usually of 1-kw carrier output and above. The limitations of the class B r-f linear amplifier are such that, for satisfactory operation, the plate power efficiency ranges from 30 to 35 per cent. Considering the driver and modulator stages and the transmitter auxiliaries, with this system the over-all efficiency from power mains to carrier power output may range from 20 to 25 per cent. The high-efficiency amplifier circuit[1] provides a plate operating efficiency as high as 60 to 65 per cent to be realized from a linear-power amplifier.

FIG. 56. Block diagram of high-efficiency power amplifier.

The amplifier circuit (Fig. 56) has been divided in block form into individual units. The exciting voltage E_x passes into two branches. One leads into a negative 90-deg phase-shifting circuit, thus transforming it to the proper amplitude for grid excitation of the *carrier amplifier* tube. This voltage E_2 is amplified by the carrier tube, the a-c components of plate voltage becoming E_1 (180 deg out of phase with E_2). The output voltage E_1 in passing through the impedance-inverting network has its phase retarded an additional 90 deg at the output of the network. Therefore, in turning through 360 deg in this path, the resultant E_0 is in phase with the exciting voltage E_x. In the lower branch of the circuit the 180-deg phase reversal of E_x in passing through the grid network and the phase reversal produced in passing through the *peak amplifier* tube result in a correct phase of E_0 at the load. The phase shifts may be further clarified by the vector diagram of Fig. 57, where the output voltages produced by both the carrier and peak tube are illustrated as acting in phase to produce E_0 at the load.

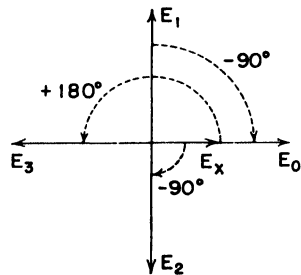



FIG. 57. Phase relations in high-efficiency amplifier.

Figure 58 illustrates the theoretical individual and combined operation of the carrier and peak tube branches of the amplifier as produced by a modulated r-f voltage (assuming sinusoidal variation of r-f voltage with modulation). For the carrier amplifier the r-f plate voltage rises very nearly linearly over the region *O* to *A*, flattening off at this point owing to saturation; beyond this point any increase in exciting voltage for this tube produces practically no further increase in plate voltage. The carrier amplifier plate current on the other hand rises quite linearly from *O* to *B*. Thus from *O* to *A* the operation is quite similar to that of a class B r-f linear amplifier operating into a load impedance of constant value, whereas from *A* to *B* there is a progressive reduction in plate impedance under influence of positive delivery of power from the peak amplifier tube on upward modulation swing, as observed through the impedance-inverting network and the plate-current rises. The plate voltage of the peak amplifier tube rises linearly from *O* to *B*, where the curve flattens off because of saturation. This tube is biased to a point where little positive power is delivered for exciting voltages below carrier amplitude *A*. However, owing to coupling to the carrier amplifier tube output circuit through the impedance-inverting network, a voltage exists in its plate circuit during this idle stage for the tube. Therefore, a linear variation in plate voltage for the carrier amplifier tube between *O* and *A* causes a corresponding linear voltage variation between *O* and *A* in the plate circuit of the peak amplifier tube because it is

[1] Doherty, W. H., A New High Efficiency Linear Amplifier for Modulated Waves, *Proc. IRE*, September, 1936.

parallel with the load. Owing to grid-biasing conditions with respect to the excitation voltage for the peak amplifier tube, appreciable plate-current flow begins when exciting voltage assumes an amplitude greater than that necessary for an unmodulated carrier condition. Over the region *A* to *B*, plate current rises very nearly linearly to the limit at *B*. It is evident that at the crest of the modulation cycle corresponding to *B* both the carrier and peak amplifier branches are delivering equal power outputs in phase to the load.

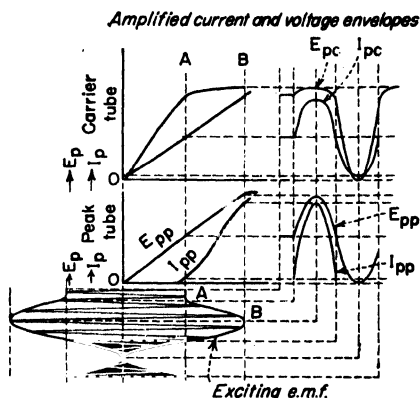
Adjustments required for satisfactory operation consist of correct neutralization of the tube capacitance, correct grid biasing of carrier and peak amplifier tubes, adjustment of grid load resistors of both amplifier tubes and their grid and output tank circuits to resonance, as well as obtaining correct phase-inverting characteristics from the circuits involved. For correct loading of the amplifier, the r-f transmission line should be properly terminated to permit operation of the amplifier into a resistive load. It will be noted that for the purpose of securing the impedance-inverting characteristics required, a 90-deg phase shift is also secured. All other phase-shift networks are utilized to compensate for this undesired phase shift. Compensation for phase shift must be effective over all useful side-band frequencies and also at the carrier frequency. The 90-deg phase-shifting circuit in the grid of the carrier amplifier tube and the 180-deg phase-shifting circuit in the grid of the peak amplifier tube are utilized for compensation purposes only.

Fig. 58. Operating characteristics of high-efficiency amplifier.

56. Ampliphase Transmitter. Another method for increasing the efficiency of modulated transmitters is due to Chireix,[1] a modern version of which is produced by RCA under the name Ampliphase. Basically the system is one in which the r-f signal is phase-modulated by the a-f signal at low level, then amplified by high-gain class C amplifiers. The modulated signal is then converted to an a-m signal in the load. Simplified diagrams of this transmitter are shown in Fig. 59*b*.

The circuit functions as follows: A crystal oscillator is fed to two independent r-f channels through phase-shifting networks so that two r-f carriers are produced out of phase with each other by 135 deg. If the output from these r-f chains were 180 deg out of phase and were impressed on a common load, no power would be developed in the load. If, however, the phase is adjusted, say by the modulating signal, so that a 90-deg phase difference exists between the two carriers, then the load current will be approximately doubled, corresponding to a modulation peak, compared with the value with 135-deg phase difference. Therefore a 22.5-deg phase shift in each r-f channel will produce full modulation.

Following the oscillator, each r-f channel consists of a stage in which the phase is brought to the 135-deg value (Fig. 59*b*) by adjusting R_1. The output of this stage, the "d-c modulator," is fed to three modulated amplifiers in each of which the phase, for full modulation, is shifted 7.5 deg. This is accomplished in a circuit similar to the d-c modulator except that the phase shift is produced by a variable-resistance modulator tube as shown in Fig. 59*b*. Conventional amplifiers and frequency multipliers bring the final power to the required values.

Because the final tubes must supply a load current that increases at modulation peaks, an increased driving voltage and increased bias are necessary to maintain the efficiency at full modulation. This is accomplished by a "drive regulator," a cathode-follower stage that samples the a-f signal, amplifies it, and applies it to the stage which drives the final.

[1] Chireix, H., High Power Outphasing Modulation, *Proc. IRE*, **23**, 1370, November, 1935. See also Starner, C. J., J. Q. Lawson, and C. D. Mulford, The RCA Ampliphase Fifty, *Broadcast News*, **84**, 54, August, 1955. See also Chap. 18, Sec. 37, of this handbook.

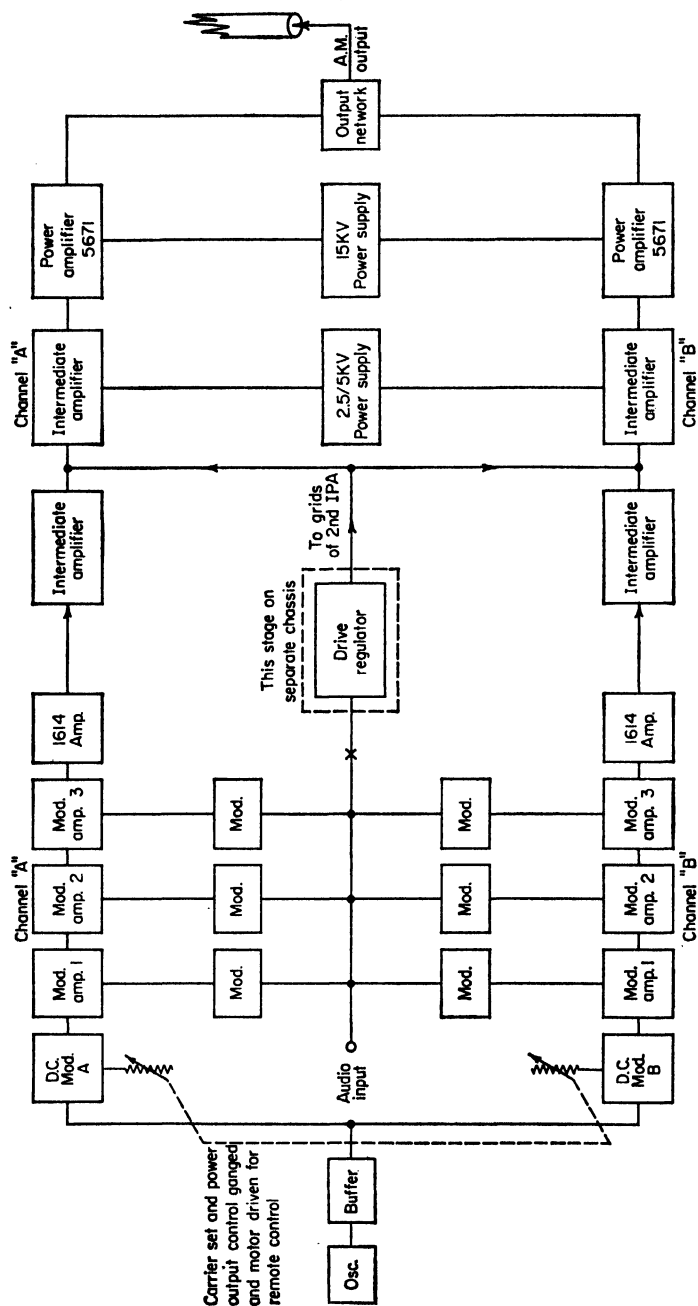



FIG. 59*a*. Block diagram of the Ampliphase transmitter. Only two high-power amplifier tubes are required compared to four in high-level design, two as r-f amplifiers and two as modulators.

Variable bias, controlled by the a-f signal, is applied to the driver to adjust the operating point. In this way about 80 per cent modulation of the driver provides 100 per cent modulation of the final.

The output circuit is a conventional pi network. About 6 db of feedback is supplied to the phase modulators to reduce the noise level and to iron out some nonlinearity in the over-all system.

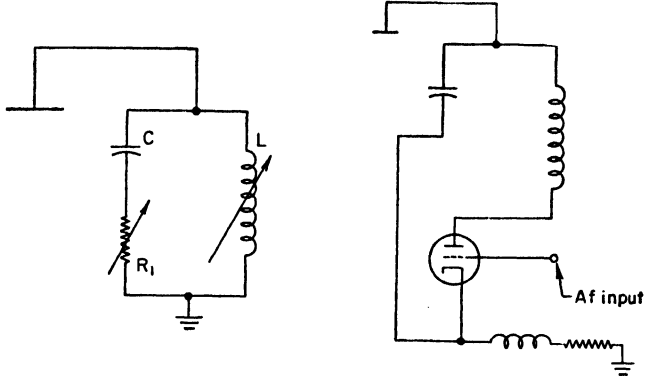



FIG. 59*b*. Simplified circuits of the Ampliphase transmitter. (*Left*) D-c modulator (r-f amplifier) in which R_1 "sets" the carrier at the 135-deg value. Low-Q tank circuit with L and C chosen so that change in R_1 changes only phase and not impedance. (*Right*) Modulated r-f amplifier which is identical to the d-c modulator except that a tube serving as a variable resistance is substituted for R_1.

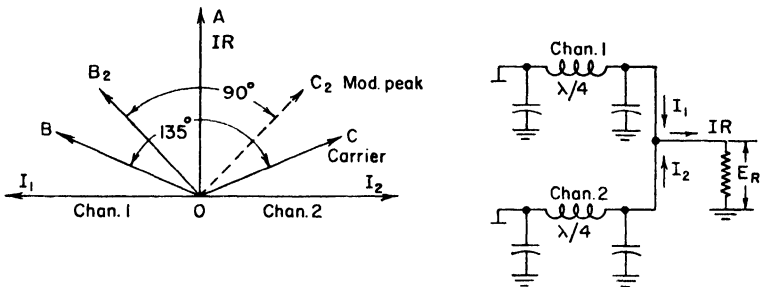



FIG. 60. Phase relations in Ampliphase transmitter.

57. Stabilized degenerative feedback as applied to broadcast transmitters reduces the harmonic distortion and noise created within the transmitter equipment, thus providing high-fidelity performance. Carrier-noise level may be as low as 65 db below 100 per cent modulation signal by utilizing degenerative feedback, even with a.c. applied to the filaments of all tubes. The FCC, Sec. 3.46, recommends that the carrier hum and extraneous noise (exclusive of microphone and studio noises) level (unweighted rss[1]) be at least 50 db below 100 per cent modulation for the frequency band of 150 to 5,000 cps and at least 40 db down outside this range. Harmonic distortion may be reduced to well below the FCC requirements, and in some cases the measured value of rms a-f harmonic distortion in the range 50 to 5,000 cps is less than 2 per cent at 85 per cent modulation and less than 3 per cent at 100 per cent modulation even with a high-efficiency power amplifier unit as a part of the system.

[1] Root sum square.

The application of feedback to the transmitter follows the same principles as applied to audio amplifiers. Application to radio transmitters is, in general, more complex than when applied to amplifiers.

In operating a transmitter with feedback the over-all gain of the audio system is reduced by the amount of feedback used. If 30 db of feedback is employed and the feedback voltage is removed suddenly by some fault, the program input will be 30 db too high, and readjustment of the program input level must be made instantaneously to prevent overloading.

With the application of feedback to cascade r-f amplifiers it becomes extremely difficult to maintain the phase of the rectified signal picked up at the output of the transmitter sufficiently close to the 180-deg rotation required throughout the entire a-f stage. Unless all networks in the entire cascade system are correctly designed, the kva/kw ratio of all r-f tank circuits is kept to a very low value, and stray capacitances are minimized, there is an accumulative phase shift through the feedback loop wherein the degenerative system is active.

Under conditions where the voltage fed back to the audio input of the transmitter after passing through the feedback loop is other than 180 deg out of phase with the input signal, less noise and distortion cancellation result. This is especially true under conditions where the phase shift of the feedback loop becomes less than 90 deg or more than 270 deg. At frequencies where the phase shift approaches zero and 360 deg from that of the input signal, stabilizing circuits are necessary to prevent violent oscillation of the entire transmitter, provided, of course, that the amplification around the loop is at least unity. These frequencies are sometimes referred to as those at which the phase "turns over." For the h-f turnover point, say around 25 kc, an adjustable stabilizing filter may be utilized in one of the low-power speech amplifier stages. This prevents oscillation of the transmitter at the particular high a.f. where the condition exists and for this reason it is called the *anti-sing* circuit. In addition there may be required an l-f (below 100 cps) stabilizing circuit in one of the low-power stages to prevent oscillation at the l.f. at which another unstable condition exists.

By correct proportioning of all constants of the a-f and r-f stages and associated networks throughout the entire section of the transmitter containing the feedback loop and by application of stabilization circuits together with careful transmitter adjustments, an effective amount of feedback can be secured for cancellation of noise and distortion.

MODULATION EQUIPMENT

58. Speech Amplifier. An audio amplifier employing power tubes is usually necessary as the preliminary part of the audio system of a transmitter to raise the audio signal to a sufficient amount to swing the grids of the modulator tubes. Resistance coupling is frequently used.

Amplitude modulation provides a means for reproducing a signal containing a distortion not exceeding a few per cent with the carrier fully modulated. In broadcasting transmitters it can be effected by either plate or grid modulation. When grid modulation is applied to a power tube, either by bias-voltage or r-f grid-voltage change, the efficiency of the power amplifier is rather low, ranging from 30 to 35 per cent. A plate-modulated radio stage operating as a class C amplifier has a comparatively high efficiency ranging from 70 to 80 per cent. This advantage of higher efficiency, however, is offset by the low efficiency of the plate modulator unless a class B audio amplifier is used for modulating. Therefore there is not much difference in the two systems, in so far as efficiency is concerned, with respect to power and vacuum-tube costs except under conditions where modulating power for a class C r-f output stage is supplied from a modulator of rather high efficiency.

When the power-amplifier stage of the transmitter is plate-modulated, the setup is called a *high-level system* of modulation, whereas a transmitter modulated in a low-power stage of the transmitter and followed by a class B r-f power amplifier is termed the *low-level system* of modulation.

59. Modulators and Modulated Amplifiers. In Fig. 61 is shown a constant-current system of modulation due to Heising.[1] The modulator and modulated amplifier are connected in parallel with a constant-current source of supply. This is connected to the common plate lead through a large inductance L_1 called the *modulation choke.*

The dynamic modulating characteristics can be determined with a fair degree of accuracy from the static characteristics of the modulator tubes in a method illustrated in Fig. 62. The modulated amplifier is assumed to be a pure resistance load in parallel

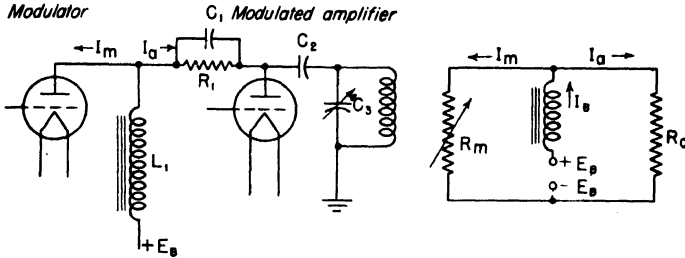



Fig. 61. Heising constant-current modulator and equivalent.

with the plate resistance of the modulator tubes and both assumed to be supplied with power through a modulation choke of infinite impedance. The sum of the instantaneous currents in the amplifier and modulator in this case is a constant. An approximation is made of the number of modulator tubes required to modulate a given r-f amplifier. The plate-current ordinate for a single tube must be multiplied by the number of modulator tubes before the load line BA can be plotted, the slope in amperes per volt which depends upon the load resistance produced by the amplifier. Line BA was chosen for two modulator tubes operating at 3,000 volts into an amplifier of 2,000 volts and 150 ma or an effective resistance of 13,333 ohms. The mean modulator plate current I_0 is chosen from allowable plate dissipation and load line BA drawn in about operating point C. The modulator grid voltage swings from the potential of the cathode to equal grid voltage on the other side of the operating point. By taking readings of plate current and voltage from end points of the load line, the following information becomes available:

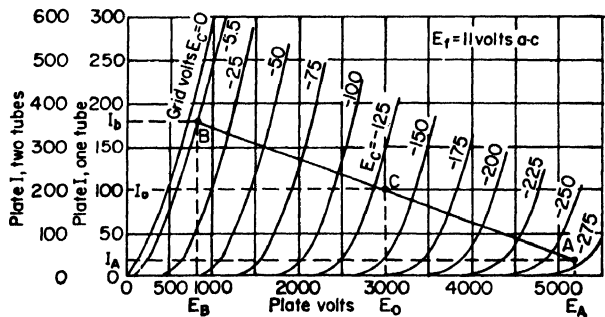



Fig. 62. Method of determining modulator characteristics.

$$\text{Modulation factor} = \frac{E_A - E_B}{2E_0}$$

where E_A = max plate-voltage swing
E_B = min plate-voltage swing
E_0 = d-c plate voltage at operating point C

$$\text{Per cent second harmonic distortion} = \frac{\frac{1}{2}(I_A + I_B) - I_0}{I_B - I_A} \times 100$$

[1] Heising, R. A., Modulation in Radio Telephony, *Proc. IRE*, **9**, 365, August, 1921.

where I_A = max plate-current swing
I_B = min plate-current swing
I_0 = plate current at operating point C

$$\text{Power output in watts} = \tfrac{1}{8}(E_A - E_B)(I_B - I_A)$$

60. Design for High Audio Fidelity. In the design of the modulated amplifier of the above system certain elements must be properly proportioned to afford a uniform frequency characteristic. C_1 (Fig. 61) should be large enough so that its impedance at the lowest frequency to be transmitted is less than one-third of R_1, or the plate-dropping resistor.

C_2 provides an r-f path from plate to filament of the amplifier tube and at the same time breaks the d-c path. It must also break the path for higher-frequency a-f current and force it to flow through the r-f amplifier tube. It should, therefore, be no larger than necessary to conduct the r-f plate current without producing excessive phase shift in the plate current under conditions where C_2 is less than $2C_3$.

Sufficient impedance of the modulation choke over the a-f range is another important factor in circuit design. Its impedance at the lowest a.f. should be at least two times the effective resistance load produced by the r-f amplifier tube. The choke should be free from inherent self-capacitance defects over the frequency range to maintain a sufficiently uniform high impedance at the higher frequencies.

High-quality signal reproduction requires that amplitude distortion should be kept at a minimum. A common cause of distortion is underexcitation of the grid of a modulated amplifier tube when plate modulation is applied. This results in insufficient driving voltage during periods of high plate-voltage swing and consequently peak-output limiting. Trouble from this cause shows up quite clearly upon an amplitude curve or upon an oscillograph in the form of chopped-off positive peaks. In Fig. 63 are shown amplitude curves taken on the modulated carrier of a stage the grid of which was excited to saturation as shown in A and underexcited in B. It is a custom to have available a surplus of driving power for a modulated amplifier to prevent any possible occurrence of amplitude distortion.

FIG. 63. Amplitude curves taken on a modulated amplifier. Curve A taken on a stage with sufficient driving power to saturate the grid and showing negligible distortion. Curve B taken on stage with insufficient grid excitation to cover the positive peaks.

The constant-current or Heising system of plate modulation is often designated as a class A system, since the modulator tube performs under conditions similar to those encountered in a class A amplifier.

A properly designed class B system permits much higher plate efficiency to be secured from a given set of tubes and correspondingly a much greater output than with a class A system. This efficiency may reach as high as 66.6 per cent with a small percentage of audio harmonic distortion.

Inasmuch as it is often necessary to drive the grids of class B audio amplifiers into their positive grid-current region to obtain maximum power output, it is important for the driver-amplifier stage for the modulator stage to have a good output voltage regulation. This calls for driver tubes having a sufficient output capacity to deliver an undistorted voltage to the grids of the class B stage, even though there is a nonuniform increase of load on the driver stage caused by the class B tubes as they are driven through the positive grid-current region of their dynamic operating characteristics.

F-M SYSTEMS

The method of program signal transmission by means of f.m. utilizes a frequency variation or deviation at the audio rate, the deviation frequency being a small percentage of the unmodulated carrier frequency.

Assume the existence of an f-m transmitter operating on 90 Mc and that a maximum deviation of ±75 kc is desired. Then a sustained sine wave of, say, 1,000 cps may be applied to the modulator audio input, with the audio signal amplitude adjusted to provide ±75 kc deviation. This would result in utilizing the full modulation capabilities of the transmitter. With a complex wave program input, the frequency deviation at any instant corresponds to the amplitude of the complex wave at that instant.

Channels for f-m transmission have been assigned 200 kc apart, which has been found to be a sufficient carrier separation to allow a frequency deviation of as much as ±75 kc. The width of the band required[1] in the frequency spectrum is *at least* twice the value of the highest modulating frequency or twice the frequency deviation, whichever is greater. Important side-band components may occur outside these limits, however.

61. F-m Methods. There are diverse methods of producing f.m. on an r-f carrier. Two rather different systems have been classified as (1) direct f.m. and (2) indirect f.m., accomplished primarily by phase modulation.

Direct f.m. is produced by frequency-modulating directly the master oscillator stage, which has a normal unmodulated carrier frequency of either the transmitter output frequency or a convenient subharmonic thereof. In papers[2] giving a mathematical treatment of f.m., it has been illustrated that, if the tank circuit constants of the oscillator stage are varied in accordance with the audio-input frequencies, there will be produced a resultant f-m output signal. With some device operating as a condenser microphone varying the capacitance of the tank circuit of the master oscillator, there may be produced an r-f carrier frequency modulated to conform with the sound undulations vibrating the microphone diaphragm. This illustrates f.m. by the direct method.

A modified form[3] of the direct system of frequency-modulating a transmitter is accomplished through the use of a tube (Fig. 64), employed as a variable reactance. Here a variable reactance is caused to exist between the cathode and anode of the reactance tube by grid-bias variation at an audio rate. By supplying the grid of the reactance tube with r-f voltage previously passed through a phase-shifting circuit C_1R_1, the grid-excitation voltage is caused to be in phase quadrature with the plate voltage. Then the a-c portion of the reactance-tube plate current i_p will be very nearly

$$i_p = g_m e_g$$

However, since

$$e_g = jKe_p$$

then, under influence of the phase-shifting network,

$$i_p = jKg_m e_p$$

and

$$Z_p = \frac{e_p}{i_p} = -\frac{j}{Kg_m}$$

from which the equivalent capacitance produced by the tube is

$$C_e = \frac{Kg_m}{2\pi f}$$

[1] Carson, J. R., Notes on the Theory of Modulation, *Proc. IRE*, February, 1922; van der Pol, B., Frequency Modulation, *Proc. IRE*, July, 1930.

[2] Roder, H., Amplitude, Phase and Frequency Modulation, *Proc. IRE*, December, 1931. Crosby, M. G., Frequency Modulation Propagation Characteristics, *Proc. IRE*, **24** (6), June, 1936.

[3] Crosby, M. G., Frequency Modulation Noise Characteristics, *Proc. IRE*, April, 1937.

The reactance tube may be caused to appear as an equivalent variable capacitance across the oscillator tank circuit.

62. Carrier Stabilization. For satisfactory transmission of the f-m signal it is essential that, in addition to producing the modulated wave, there must be a satisfactory carrier-wave stability. For this reason it is necessary to add a stabilizing circuit by which the average frequency of the carrier wave is compared to that of a precision crystal oscillator. The resulting difference frequency when applied to a special motor coupled to a tuning compensator provides carrier-frequency correction to the assigned channel frequency.

In Fig. 65 are shown the basic circuits for monitoring the transmitted carrier frequency within the required frequency tolerance. The system consists of two balanced

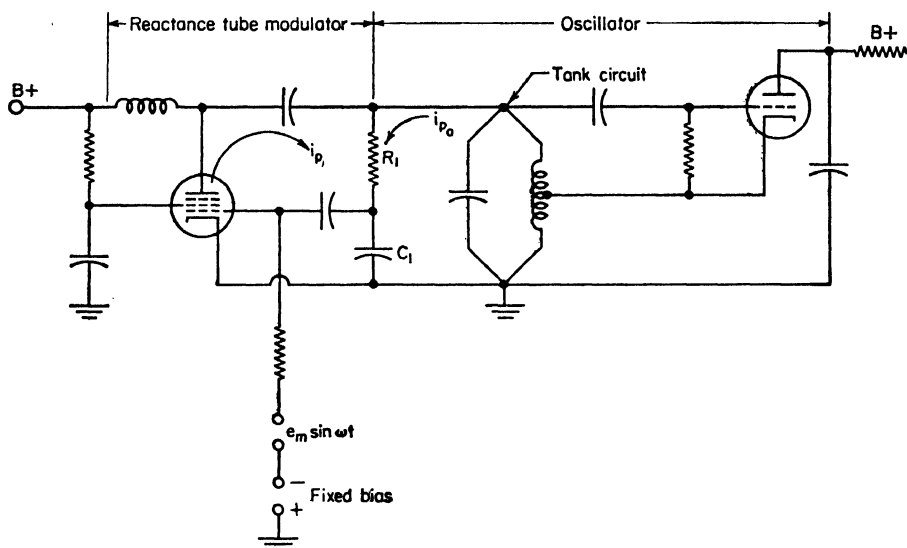



FIG. 64. Reactance-tube modulator.

modulators arranged to secure in the output of each a beat frequency equivalent to the difference between a crystal-frequency standard and that of the master oscillator. The output of the crystal oscillator is divided to feed equal amounts of energy to the two balanced modulators. Two phase-shifting networks provide a 90-deg phase displacement between the input signals applied to the two balanced modulators.

Since the signal from the master oscillator provides a signal of the same phase for both modulators, then the 90-deg phase displacement in crystal-oscillator output signals supplied to the modulators will provide a similar phase displacement in the beat-frequency output from the two modulators. In this manner two-phase energy is available for driving a tuning motor to compensate frequency shift of the master oscillator. The direction of rotation of the 90-deg vectors representing the signals at the outputs of the two balanced modulators changes from clockwise to counterclockwise depending upon whether the master-oscillator frequency is higher or lower than the crystal standard frequency.

The two-phase a-c output of the two modulators, therefore, has the characteristic of reversing one phase when the output secured from the master oscillator swings through zero beat from a higher to a lower frequency compared to the crystal.

The two-phase a-c output of the two modulators thus secured may be used to energize the windings of a two-phase motor. If the master-oscillator frequency is higher than that of the crystal, the motor shaft will rotate clockwise and so on. A com-

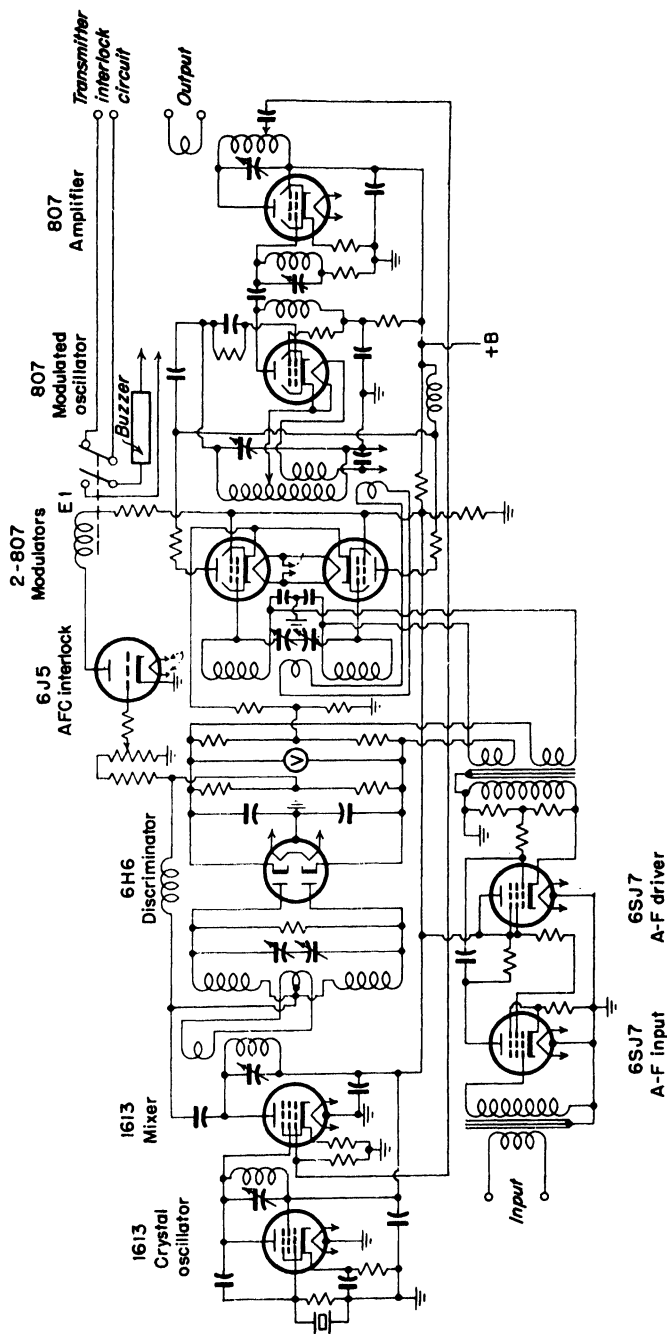



Fig. 65. Complete f-m transmitter circuit producing a compensating voltage from a frequency deviation.

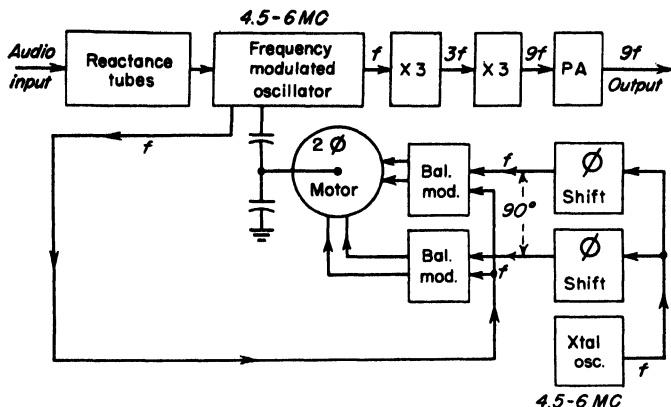



FIG. 66. Frequency-control f-m exciter.

pensating variable tuning capacitor may be attached to the motor shaft and employed to retune the master oscillator to zero beat with the crystal, which operates on a submultiple of the assigned carrier frequency.

Elements of the RCA BTF-3B reactance-tube modulator f-m system are shown in Fig. 66. The operating characteristics of this transmitter follow:

Frequency range	Any specified frequency between 88 and 108 Mc
Power output (into transmission line)	1,000-3,000 watts
R-f output impedance	35-75 ohms
Carrier-frequency stability, deviation less than	2,000 cps
Modulation capability	±100 kc
Method of modulation	Reactance tubes
Audio input impedance	600 ohms
Avg program level	+4 ± 2 vu
100% modulation level	+12 ± 2 dbm
A-f response*	
30-15,000 cps, uniform within	±1 db
A-f distortion†	
30-100 cps	1.5%
100-7,500 cps	1.0%
7,500-15,000 cps including all harmonics up to 30 kc at 75-kc swing	1.5%
F-m noise level, below ±75-kc swing†	65 db
A-m noise level, below 100% amplitude modulation†	50 db

* For preemphasized response the preemphasis filter is provided to be inserted in the 600-ohm audio-input line at the most effective point.

† Distortion and noise are measured following a standard deemphasis network.

The 1-kw Western Electric type 503B-2 f-m[1] transmitter utilizes reactance tubes directly as frequency modulators in a manner as shown in Fig. 67. The method used to maintain constant the mean carrier frequency is called "synchronous f.m." since it operates by comparing the mean frequency (measured in total carrier cps) of the f-m oscillator to a precision fixed-frequency standard. The difference frequency thus derived is then utilized as a control medium for mechanically retuning the oscillator stage, keeping the oscillator frequency an exact multiple of the standard.

Figure 68 shows the frequency-stabilizing system, which functions through a small portion of the 6-Mc f-m oscillator (assume a 96-Mc carrier), being fed back through frequency dividers to obtain a 6-kc frequency equal to that of the precision quartz-crystal frequency standard. The 6-kc frequency, a much lower submultiple of the 96-Mc carrier, is necessary to produce a difference frequency sufficiently low to be within the range of the rotating magnetic field of the electric motor used for retuning

[1] Ganzenhuber, J. H., Western Electric FM Broadcast Transmitters, *FM & Television*, September and October, 1946.

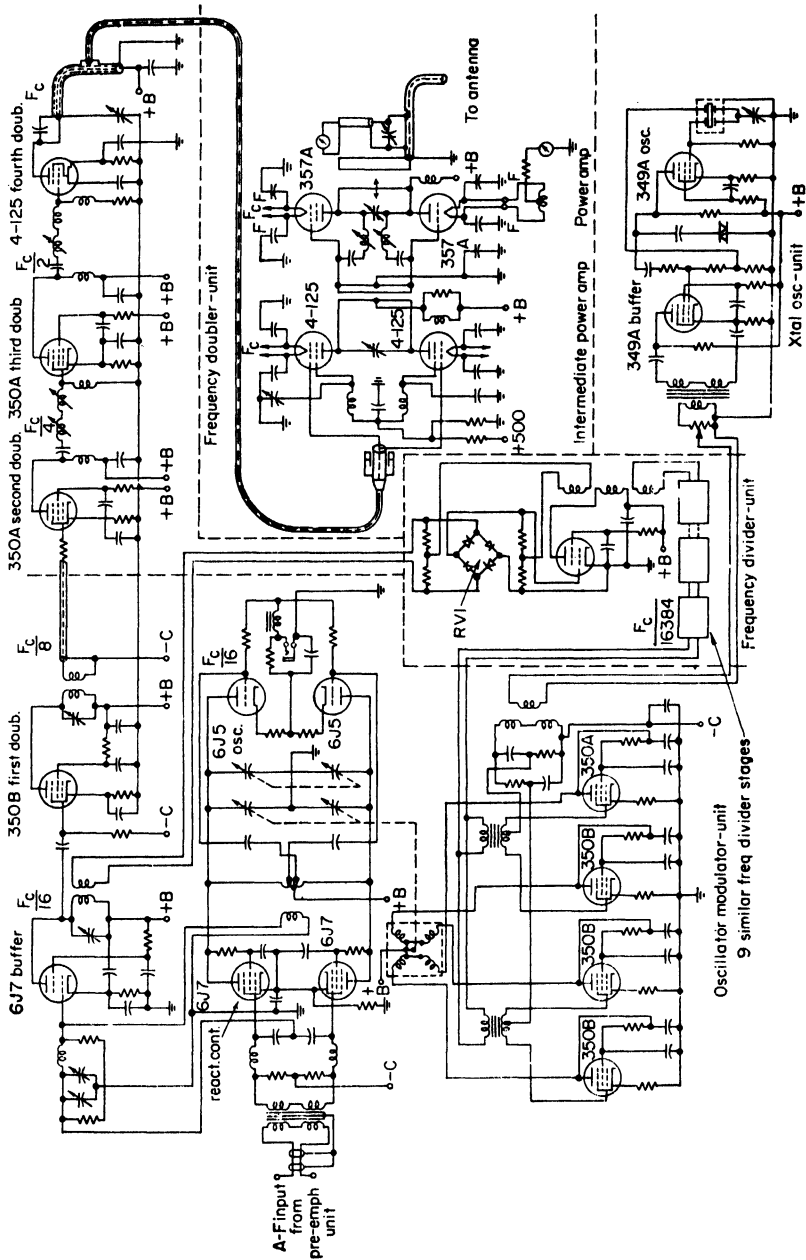



FIG. 67. Essential circuits of the 1-kw Western Electric f-m transmitter.

the oscillator variable capacitors through a suitable speed-reduction mechanism of gear trains. The direction of rotation of the motor depends upon whether the oscillator frequency is higher or lower than the fixed multiple frequency of the standard. Thus automatic readjustment of the oscillator tuning is made in the correct direction so that when exact synchronism occurs between the frequency fed back from the f-m oscillator and the frequency standard the motor is at rest.

Because of the inertia of the motor rotating elements and the high order of frequency division used, the motor is not caused to rotate by frequency deviations produced on

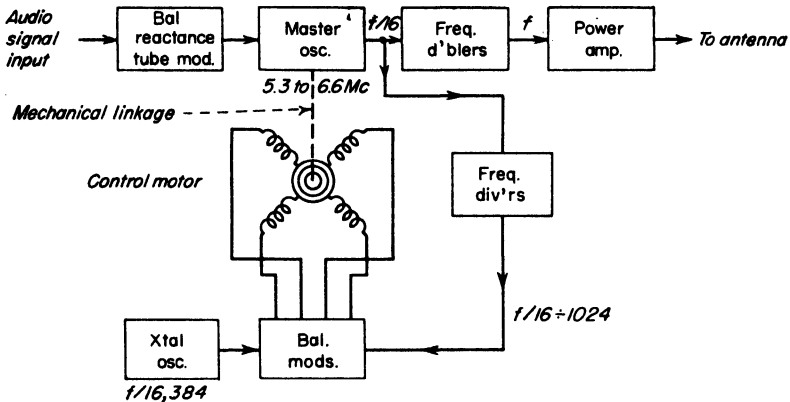



FIG. 68. Frequency-stabilizing system of Western Electric.

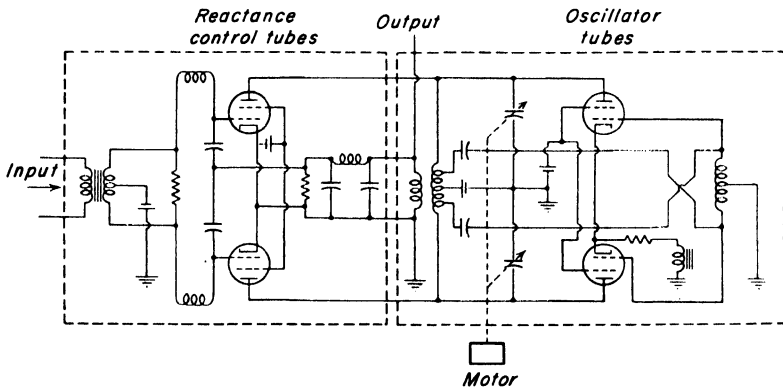



FIG. 69. Reactance-tube circuit of Western Electric f-m transmitter.

the carrier at an audio rate during modulation. The main advantage claimed for the synchronized f-m method is that the output frequency is maintained identical in precision to the standard by making all the controlling factors in terms of frequency.

63. Armstrong F-m Method. The indirect or p-m method of producing f.m.[1] consists in general of a constant-frequency oscillator, a modulator (the function of which is to change the phase of the oscillator output as illustrated in Fig. 70), and a series of multipliers to increase the amount of p.m. sufficiently to secure the frequency shift or modulation required in the radiated signal. Results are secured by splitting the oscillator output into two paths. One path contains a phase-shifting network which

[1] Armstrong, E. H., A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation, *Proc. IRE*, May, 1936.

shifts the phase 90 deg and the other a balanced modulator generating side bands with a suppressed carrier.

A combination of these two signals produces a p-m signal with a phase-shift modulation capability up to ±30 deg with satisfactory linearity. An f-m wave is derived therefrom by transmitting the signal through frequency multipliers. A multiplication of several thousand times is required to obtain deviations of ±75 kc.

To produce f.m. and at the same time maintain a constant deviation frequency, the p.m. must necessarily be inversely proportional to the modulating frequency. Therefore with this method it becomes necessary to have the amplitude of the p-m signal decrease in proportion to the frequency of the audio input to secure a flat a-f transmitter response. This is usually accomplished by a corrective network in the audio circuits ahead of the modulator.

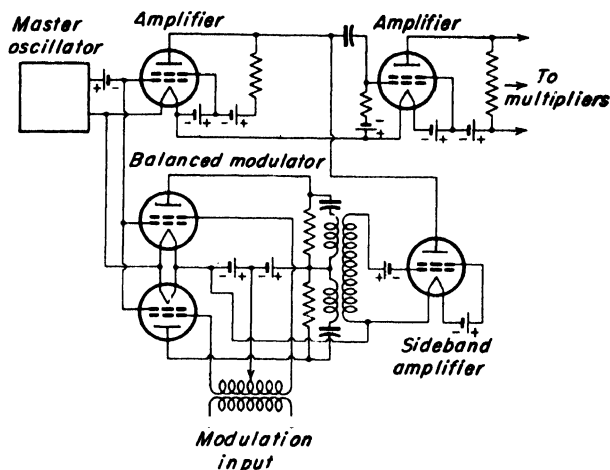



FIG. 70. Armstrong frequency modulator.

The amount of frequency multiplication required following the production of phase modulation to secure the desired f.m. by the indirect method depends upon (1) the amount of p.m. produced by the modulator, (2) the lowest a.f. transmitted, and (3) the deviation or frequency swing at the output frequency. For a phase shift of 0.5 radian, frequency deviation of ± 60 kc, and lowest a.f. of 30 cps, the frequency multiplication required is 4,000 times.

To obtain this amount of frequency multiplication, the initial oscillator frequency must be multiplied in several stages, then heterodyned down to a lower frequency, and again multiplied a number of times to secure the output frequency.

64. Merits of F.M. vs. A.M. With the application of f.m. to transmitters operating in the v-h-f band, the relative merit for this system of signal transmission can be evaluated on the basis of an a-m system. The v-h-f signal field intensity[1] at a given distance from a particular transmitting antenna may be determined from theoretical and empirical relationships[2] and derived from extensive mathematical and experimental work. Actual experimental tests[3] have shown that an interfering audio signal (output of receiver) will create objectionable interference if its level is about 30 to 40 db below the desired signal. Thus service areas can be defined as zones in which the desired component of the resulting audio signal is at least 35 db above

[1] Trevor and Carter, Notes on Propagation of Waves below Ten Meters, *Proc. IRE*, March, 1933.

[2] Devino and Hunt, Ultra Short Wave Propagation over Land Burrows, *Proc. IRE*, December, 1935

[3] Weir, I. R., Field Tests of Frequency and Amplitude Modulation with U-h-f Waves, *Gen. Elec. Rev*, May, 1939; Crosby, M. G., The Service Range of Frequency Modulation, *RCA Rev.*, January, 1940; Guy, R. F., and R. M. Morris, NBC Frequency Modulation Field Tests, *RCA Rev.*, October, 1940.

the interference. For very-high-quality reproduction, this figure runs from 40 to 55 db.

For interfering signals on the same channel as the desired signal, it is evident that, if a.m. is used, a signal input ratio of 35 db is required to secure the desired output ratio. However, in f.m. the ratio of signals at the receiver input needs to be only about 6 db since the receiver for f-m reception responds to frequency variations and limits amplitude variations such as those caused by noise and undesired signals.

On this basis there are advantages of f.m. over a.m. because of (1) improved signal-plus-noise to noise ratio. Experimental results have shown this difference to be as much as 25 db as influenced by intensities of automobile ignition, X rays, and other man-made interference. Atmospheric interference being small at v.h.f., it becomes negligible in comparison with man-made interference. (2) A uniform and definite service area from a given transmitter since f-m signal-plus-noise to noise ratio remains high until field intensity reaches a low value. (3) A smaller geographical interference area obtained when two f-m transmitters are operated simultaneously on the same frequency as compared to similar operation of two a-m transmitters. (4) An r-f amplifier used to increase an f-m signal is more efficient than one used for a.m. because f.m. can be accomplished at low level followed by a class C r-f power amplifier. (5) For a given service area, less radiated power is required for f.m. because of the improvement in signal-plus-noise to noise ratio obtained with f.m. (6) For a given power output, power-tube operating costs are less because relatively smaller tubes can be used for f.m. for a given power output.

The FCC has provided for 65 channels each 200 kc wide for f-m carrier frequency assignments between 88 and 108 Mc.

Preemphasis. A-f preemphasis at the transmitter is added as a means of increasing the effective modulation at the higher audio frequencies since, for average programs, the peak levels of modulation at these frequencies are lower than for the lower frequencies. At the receiver a corresponding amount of deemphasis is employed to bring the modulation back to its proper level. In this manner an increase in the signal-to-noise ratio is brought about.

Preemphasis is effective only when the inherent noise level of the transmitter at the point of preemphasis is low, since any noise or higher frequency disturbances already existing would be emphasized along with the microphone modulation. Since f-m transmitters generally have less noise (as required by the FCC) than a-m transmitters, this form of enhancing the higher audio frequencies is more effective with f.m. than with a.m.

Preemphasis is generally expressed in terms of the time constant of an *LR* circuit which produces the effect. The required value is that corresponding to a time constant of 75 μsec.

An important advantage of the f-m transmitter over one utilizing the amplitude system lies in the fact that the efficiency of the r-f amplifier stages can be as great as it is for class C telegraph service. At v.h.f. this may be between 50 and 70 per cent.

Transmitter Measurements. The measuring equipment is considerably different than is required for an a-m station since there is a variation in frequency of the emitted wave with modulation while the amplitude is kept more or less constant. This is exactly the reverse of a.m., where the carrier is varied in amplitude but maintained at a constant frequency. The frequency swing or deviation can be measured by applying sustained tone to the transmitter and then measuring the relative intensities of the carrier and the side frequencies present, the relative amplitudes of which correspond to the Bessel functions involved.[1]

TRANSMITTER DETAILS

65. Air- and Water-cooled Tubes. For tubes of low power, artificial cooling during operation is usually not necessary, radiation into the air being sufficient. For the larger tubes, however, artificial cooling is usually accomplished by means of a circulat-

[1] Crosby, M. G., A Method of Measuring Frequency Deviation, *RCA Rev.*, April, 1940.

ing-water system which causes a sheet of water to pass over the anode surface at very high velocity.

To restrict leakage of current from the anodes to the grounded pipes of the water system, connection is made between the anodes and the water system through a long length of coiled hose or porcelain tubing. This interposes, between the anode and ground, columns of water long enough to make the electrical resistance to ground very high; as much as 100 ft of coiled hose may be used, giving resistances of 0.5 up to several megohms.

In many cases distilled water is used, the water being maintained at a satisfactory temperature by an artificial cooler, since for economical reasons it is desirable that the same water be used indefinitely.

The water-cooling and circulating system is automatically started when the transmitter is turned on, and the transmitter is automatically turned off in the event of any failure in the water-cooling system. One method of doing this employs a Venturi tube whose inlet and output orifices are connected to a device operated by the difference in pressure established between the two orifices by the flow of water. If the flow is interrupted or falls below its normal value, a contactor through additional relays causes the power supply to be disconnected.

Sometimes a milliammeter is provided on the transmitter panel which indicates the magnitude of the current leaking through one of the closed coils, the amount of current serving to indicate the relative purity of the water and indicating when it is advisable to change the water supply.

In place of water cooling, forced-air cooling is also used on some large tubes. For the large dissipation required, a large number of radiating fins are made a part of a copper radiator attached to the copper anode. Sufficient air is forced upward and between the cooling fins to carry away the heat developed on the anode. Because of the high electrostatic capacitance created by these anodes, they are not used on the very high frequencies with transmitters of very high power.

66. Power Supply. Plate-voltage supply for transmitters may be obtained from d-c generators, high-vacuum tube rectifiers, mercury-arc rectifiers, or hot-cathode mercury-vapor rectifiers.

Table 2. Comparison of High-vacuum and Mercury-vapor Tube Rectifiers*

No. of tubes	Tube type	Circuit	D-c output			Tube drop		Losses, kw		Efficiency, per cent
			Volts	Amp	Kw	Volts	At amp	Filament	Tube drop	
6	UV-214	3ϕ, double-Y	15,000	12	180	1,560	6	6.9	18.7	85.7
6	UV-857	3ϕ, full-wave	15,000	12	180	15	12	1.5	0.36	98.8
6†	UV-857	3ϕ, full-wave	21,000	30	630	15	30	1.5	0.9	99.4

* *Proc. IRE*, **18** (1), January, 1930.
† Maximum rating.

The hot-cathode mercury-vapor rectifier is considered the best method of supplying high voltages to transmitter plate circuits. The most striking difference between mercury-vapor tubes and high-vacuum tubes is the internal voltage drop between plate and cathode. In the high-vacuum tube the voltage drop may vary from a few volts to several thousand volts, depending upon the current, element spacing, etc. In the mercury-vapor tube the space charge is limited by the arc drop of the vapor which is practically constant at values between 12 and 17 volts regardless of the current.

Table 2 gives a direct comparison of the relative efficiencies of a high-vacuum tube and a mercury-vapor tube. Note that the mercury-vapor tubes give very low internal voltage drop and have considerably higher efficiencies.

Two fundamental limits determine the power output that can be obtained from any number of tubes operated in any type of circuit: (1) the maximum peak inverse voltage at which the tube can operate without flashing back and (2) the maximum peak plate current which the cathode can supply with a reasonably long life.

The maximum peak inverse voltage which can exist across a tube in any of the usual types of circuits is equal to the line-to-line peak or crest voltage of the power transformer less the voltage drop of the conducting tube.

The peak plate current depends upon the type of circuit, tube, filter, and load. In a single-phase full-wave circuit each tube must carry the full-load current for half the time. In the three-phase half- and full-wave circuit each tube carries the load current for one-third of the time. If the rectifier feeds into an inductance, square blocks of current are drawn from the rectifier and the peak plate current approaches the d-c value. If the rectifier feeds into a capacitance load, plate current is drawn for only a part of each half cycle and the peak current may reach values of from three to five times that of the d-c load current.

Table 3. Hot-cathode Mercury-vapor Tube Ratings

Tube type	Filament		Peak inverse voltage	Peak anode current, amp
	Volts	Amp		
866A	2.5	5	10,000	1.0
872A	5	7.5	10,000	5.0
869B	5	18	20,000	10.0
857B	5	30	22,000	40.0

Table 3 gives data on several typical hot-cathode mercury-vapor tubes designed for radio-power-supply purposes. Single-phase full-wave and three-phase half-wave circuits are quite generally used. The three-phase full-wave circuit is particularly applicable to the half-wave mercury-vapor tube, since it gives a peak inverse voltage whose magnitude is only 4.5 per cent greater than the average output voltage; the wave form is that of a six-phase rectifier.

FIG. 71. Class B r-f amplifier arranged for parasitic suppression.

67. Parasitic Oscillations. One of the most important design features of a transmitter is to provide for adequate suppression of parasitic oscillations. Such spurious oscillations are usually caused by regeneration in an amplifier stage. They have frequencies different from the fundamental or its harmonics.

All classes of amplifiers are subject to these oscillations. Suppressing them in a class C amplifier is not usually so difficult as in the class B types, where the grids of the tubes are driven positive for a considerable portion of the cycle. Before reliable and economical service can be realized from a transmitter of any type, all tendencies for parasitic oscillation must be suppressed to prevent serious lessening in the life of vacuum tubes or program interruptions because of arc-overs in the transmitter. Such oscillations may exist in an otherwise normal amplifier stage and may not be evident to casual inspection owing to their disappearance entirely when grid excitation is removed.

A typical class B r-f power-amplifier stage is shown in Fig. 71. This amplifier con-

tains inherent design features which have a tendency to suppress spurious oscillations. C_6 and C_7 assist by acting as a very low reactance path for all parasitics of a frequency higher than the fundamental with a result that they effectively load the parasitic circuit. Connections between these capacitors and the tube grids are kept at an absolute minimum. Loading resistors R_1 and R_2, the real purpose of which is to improve the regulation of the grid circuit as the grids swing positive, also act as a resistor load to damp out oscillations. C_8 and C_9, with their mid-point grounded, act as low-reactance paths to ground for frequencies above the fundamental.

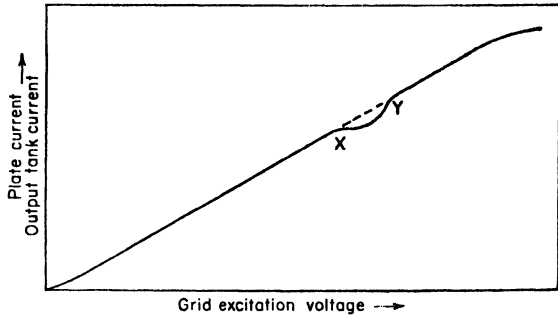
The frequency of parasitic oscillations may be anything from the very low end of the frequency spectrum to the u-h-f region. Parasitics of very low frequencies, in the neighborhood of less than 1 to 10 cps, are sometimes set up by the dynatron action of the tubes at the natural period of the power-supply filter circuit C_1, C_2, and L_1.

FIG. 72. Typical saturation curve of class B r-f linear amplifier showing dynatron effect of power-tube grids with E_b and E_c constant.

The existence of these low-frequency parasitics usually becomes apparent in the form of a severe irregularity in the saturation curve of the linear amplifier. Such a curve is shown in Fig. 72. The point X shows the beginning of this parasitic condition and Y the point where it ceases. It is caused by the dynatron characteristics of the amplifier tube grids and occurs at a point on their operating characteristic just before they are driven positive. A solution for such a condition is to use tubes whose amplification factor is such that the region XY falls below the carrier operating point. For this reason high-mu tubes have on some occasions been found to be more satisfactory than low-mu tubes.

FIG. 73. Equivalent parasitic circuit of Fig. 71.

Low-frequency oscillations of approximately one-third to one-fifth of the fundamental frequency are sometimes caused by tuned-grid tuned-plate regeneration with the plate chokes L_5 and L_4 in combination with C_{10} and C_{11} forming an output tank circuit. A similar grid tank circuit is formed by C_6, C_7, and L_3. Inasmuch as all tubes are effectively in parallel for this combination, the neutralizing capacitors tend to aggravate the condition rather than to prevent it. In Fig. 73 is shown an equivalent parasitic circuit of the combination as formed from the circuit in Fig. 71. The remedy is to change the values of L and C in either the parasitic grid or plate circuits so as to cause their natural periods to depart substantially from a near resonance condition. It is usually possible to suppress such oscillation by tuning the parasitic grid circuit to a higher frequency than the corresponding plate circuit.

The existence of these oscillations may usually be detected by applying excitation at the fundamental frequency to a stage with reduced plate voltage and grid-bias voltage until the tubes draw plate current. If oscillation of the stage continues after

fundamental grid excitation is removed, as indicated by neon lamps attached to the tube plates, the frequency of the parasitic may be determined by means of a wavemeter, and thus steps can be taken to eliminate it.

Oscillations within an amplifier stage at frequencies near the fundamental are usually caused by regeneration within an amplifier stage due to improper neutralization causing tuned-grid tuned-plate circuit oscillations. Improper circuit design or too close coupling between the inductances of the input and output circuits or chokes is also liable to cause this condition.

Parasitics of frequencies in the neighborhood of from 5 to 20 times the fundamental result in cases where the leads from the tube grids and C_6 and C_7 form a grid tank circuit, the resonance frequency of which is determined by various distributed capacitances and the inductance of the leads. Oscillations are made possible by the existence of a similar plate tank circuit formed by leads from the tube plates to C_8 and C_9 together with various stray capacitances. This form of parasitic is seldom sustained but shows itself most prominently when the stage is subject to high peaks of modulation. The trouble may usually be corrected by choke coils L_7 and L_8 inserted in the plate leads at a point adjacent to the tube plates.

These coils L_7 and L_8 together with a shortening of grid leads to an absolute minimum may also assist in suppressing oscillations of ultrahigh frequencies in amplifier stages employing two tubes in parallel. The grid leads of the two tubes, although connected, may combine with stray capacitance, thus forming a push-pull v-h-f oscillation. Such oscillations cause high r-f voltages to build up which may result in serious arc-overs from various parts of the tube output circuits.

68. Suppression of R-f Harmonics. It is the inherent characteristic of a vacuum tube, while functioning at a reasonably high efficiency as an amplifier, to generate harmonic frequencies of the fundamental. A station on 600 kc, if second and third harmonics were not suppressed, would produce interference with other stations on 1,200 and 1,800 kc. Field intensity measurements about a station are necessary to determine how much harmonic energy is radiated and to show the progress of work done toward reducing radiation.

In specifying the allowable harmonic radiation from a broadcasting station the IRE Committee on Broadcasting as of January, 1930, recommended that the maximum radio field intensity of a harmonic component measured at a distance of 1 mile from a station should not exceed 0.05 per cent of the field intensity of the fundamental. The FCC regulations prohibit radiation of a spurious or harmonic nature which may cause interference.

A field strength of, say, 500 μv per m at a distance of 1 mile was recommended as a maximum allowable intensity from a high-powered station. If in the case of a 50-kw station a circular-field pattern and equal attenuation are assumed for both a harmonic and the fundamental in the immediate vicinity of the station, a field strength of 500 μv at 1 mile would correspond to approximately 7 mw of radiated power at a harmonic frequency. The effect of directivity (illustrated in curve B, Fig. 74) may cause a field intensity of a number of times the value of 500 μv to be projected in a given direction with a very small fraction of 1 watt of harmonic power in the transmission-line and antenna circuits. Such a concentration of radiated power may form very objectionable interference. Considering the factors involved, therefore, it is evident that harmonic suppression must be attacked from a number of angles. These may be briefly outlined as follows:

1. Design of the transmitter circuits to reduce the harmonic content of the power delivered to the antenna circuits to a minimum.
2. Thorough and effective shielding of the entire transmitter or building.
3. Effectively grounding all harmonic drain circuits and elimination of long conductors near the transmitter coupled to it inductively or capacitively.
4. Reduction of directivity of harmonic radiation to a minimum.
5. Installation of shielded band- or low-pass filters at the input end of the transmission line to the antenna.

6. **Installation of an electrostatic shield between the power output stage and the antenna coupling circuits.**

Some commonly used triode circuits are shown in Figs. 75 and 76. The push-pull amplifier is superior to the single-ended circuit, as it is capable of producing a sum plate current of the two tubes which is symmetrical in wave shape and, therefore, it contains no even harmonics. Individual plate currents, of course, contain even

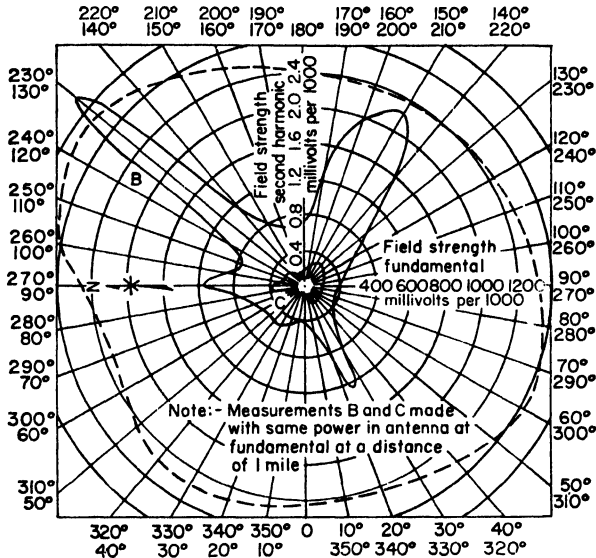



FIG. 74. Radio field-intensity survey. Dashed curve gives fundamental frequency field strength; B and C are second-harmonic intensity before and after reduction.

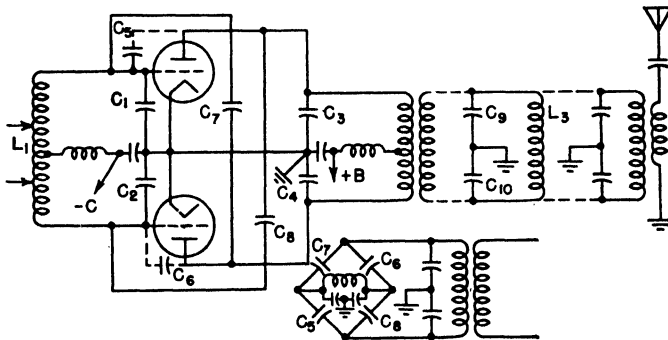



FIG. 75. Push-pull amplifier with high-kva tank circuit in transmission line.

harmonics which are drained to ground through C_3 and C_4, resulting in identical instantaneous even-harmonic potentials on each side of the tank primary but no actual even-harmonic current through it. Under these conditions an electrostatically shielded inductive coupling is provided to permit transfer of only fundamental and odd harmonics to the coupled circuit. For a condition of symmetrical plate current the tube characteristics must match closely, $C_1 = C_2$ and $C_3 = C_4$. The neutralizing bridge

must be balanced not only for the fundamental frequency but for even harmonics. This requires that the internal capacitances of the tubes should match. As will be shown later, a high ratio of circulating kilovolt-amperes in the tank circuit to the kilowatts delivered from the amplifier reduces the output of harmonics from a single-ended amplifier to a very low value. This is also true in the push-pull circuit.

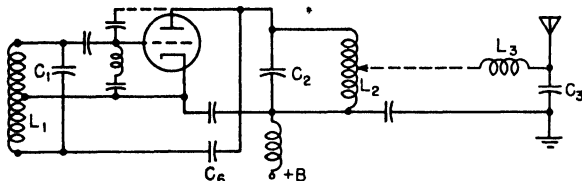



FIG. 76. Line termination effecting reduced harmonic radiation.

The circuit shown in Fig. 76 will give a very small amount of harmonic output by proper design of the circuit constants. The curves in Fig. 78 show the filtering effect of a high kilovolt-ampere tank circuit in suppressing harmonic components of current generated in the tube. These curves show actual harmonic transferred to a given load circuit Z_L with a constant output at the fundamental and various kilovolt-ampere-to-kilowatt ratios of L_2 and C_2. Figure 77 shows improvement in tank circuits so as to increase the normal filtering action of an ordinary tank circuit. A high kva/kw ratio applied to these circuits is capable of reducing harmonic output to an extremely small amount. There are some limitations to the amount of filtering which can be secured by a high-kva tank circuit, however, since the I^2R losses in the circuit increase in proportion to the circulating kva and the cost of apparatus for increasing kilovolt-amperes in a circuit without increasing losses is considerable. In broadcasting transmitters there is the limitation of too low a decrement in a circuit attenuating too greatly the high frequencies of a modulated envelope. In Fig. 77*b* L_3C_3 is tuned to a particular harmonic to be eliminated. The use of antiresonant circuits (parallel traps) in the plate lead of an amplifier, while reducing to some extent a single harmonic, has a tendency to allow considerable voltage to build up at others. Most satisfactory results are usually secured by designing a minimum-impedance path for harmonics to ground as compared with a given high impedance at the fundamental.

FIG. 77. Improved tank circuits for suppressing harmonics.

FIG. 78. Effectiveness of high kva/kw ratio in reducing harmonic output with constant power output at the fundamental.

The effectiveness of the shielding may be determined by operating the transmitter

with full power output into a shielded phantom antenna. Measurement of the harmonic field strengths produced from the transmitter itself is direct evidence of how well it is shielded. Such radiation can usually be traced to a long conductor near the transmitter, coupled to it capacitively or through a common ground return. Ground conductors serving to drain harmonic frequency power to ground therefore should be as direct as possible and should not be extended so as to have a free end which might attain a high potential at resonant frequencies. This is particularly true of the harmonic drains near the antenna itself. These should have a separate ground to prevent coupling of harmonic frequencies into the antenna.

A sensitive wavemeter is very useful in determining the relative harmonic field intensities near the various circuits of a transmitter. When tuned to the frequencies of various harmonics and coupled to various circuits of the transmitter or placed at positions along nearby open conductors, this instrument will indicate proportionate amounts of the harmonic components of the current flow. By effectively grounding a long open conductor, either directly or through large capacitors at a number of distributed points, harmonic radiation can usually be eliminated.

The push-pull amplifier coupled to a long transmission line has often become a source of undesirable even-harmonic radiation because of sufficient electrostatic capacitance existing between the coupled circuits to permit a transfer of energy from the amplifier output circuit to the line. Unless this capacitance is reduced to an extremely low value, *e.g.*, by installation of a well-grounded electrostatic screen between the two coils, even harmonics usually find a path along the transmission line with a ground return to the generating source. An unshielded transmission line serves in this case as an effective directive radiator in the form of a large loop. Its effective height will be dependent upon the height of the transmission line above ground. Parallel flow of even-harmonic currents along the line, therefore, makes it a much more effective radiator in some directions than the push-pull flow of harmonic currents in the line.

A circuit very effective in reducing both the parallel as well as the push-pull flow of harmonic currents in a transmission line is shown in Fig. 75 in the form of a high-kva floating tank circuit $L_3C_9C_{10}$ tuned to the fundamental component of current flowing in the line. This tank circuit, while offering an impedance to the fundamental approaching an infinitely high value, offers a relatively low-impedance path to ground for the parallel flow of even harmonics equivalent to

$$Z_{ne} = \frac{-1}{4\pi f_{ne}C_9} = \frac{-1}{4\pi_{ne}C_{10}}$$

where resistance of circuit is negligible, Z_{ne} = impedance to nth even harmonic, and f_{ne} = frequency of nth even harmonic, and for the push-pull flow of odd harmonics between transmission-line conductors

$$Z_{no} = \frac{-2\pi f_{no}L_3}{(2\pi f_{no})^2L_3C - 1}$$

where resistance of circuit is negligible, Z_{no} = impedance to nth odd harmonic, f_{no} = frequency of nth odd harmonic, and $C = C_9/2 = C_{10}/2$, where $C_9 = C_{10}$.

As C_9 and C_{10} are increased in value the effectiveness of the circuit in reducing harmonics is increased. Since the transmission-line termination impedance is usually made to match the line impedance for the fundamental frequency, it usually happens that the line impedance is matched for this frequency only and as a result harmonic components of current and voltage in the line appear as standing waves along the line. In such a case the above tank circuit is most effective for eliminating a particular harmonic if it is placed at a point along the line of maximum voltage. This circuit alone was effective in one case in reducing second-harmonic radiation from a station to one-fifth of its former value.

Antiresonant circuits installed in a transmission line at current antinodes have been found very effective in reducing a single harmonic to which they were tuned. Extreme

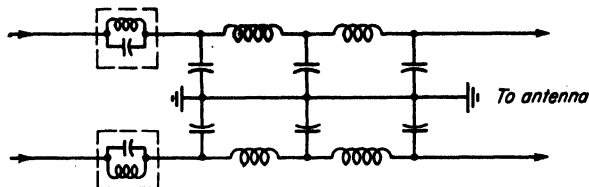



FIG. 79. Low-pass r-f filter combined with antiresonant circuits in transmission line.

care should be taken in shielding these antiresonant circuits to secure best results. A combination of antiresonant circuits and a low-pass filter is shown in Fig. 79. This combination has been used successfully in severe cases of harmonic radiation from a very long transmission-line and antenna system. The filter matches the surge impedance of the line and has a cutoff frequency between the fundamental and

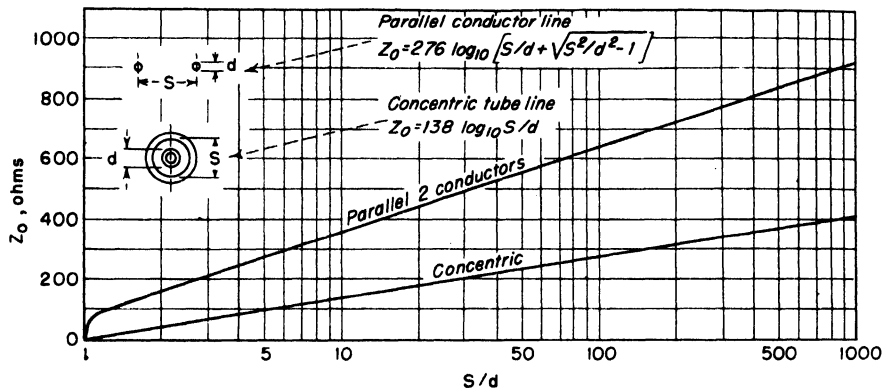



FIG. 80. Impedance of parallel and concentric-tube lines.

second harmonic. Antiresonant circuits have been found useful to sharpen the cutoff so as to attenuate sufficiently the second-harmonic frequency. Considerable experience in filter design and adjustment is required to secure optimum results from such an arrangement. For use with concentric lines with the outer sheath grounded, the filter shown in Fig. 79 is simplified to the extent of one-half, *i.e.*, one line to ground.

Methods of line termination shown in Figs. 75 and 76 are effective in reducing the possibility of harmonics reaching the antenna circuit. The termination shown in Fig. 76 may be improved by use of a multisection low-pass filter.

69. Antenna Circuit Termination for R-f Lines. Considerable improvement in antenna efficiency can be secured from an antenna located at some distance from the station so as to approach the ideal case of an antenna radiating in free space. An r-f transmission line is used for conveying the energy from the transmitter to the antenna. A simple form of such a line is

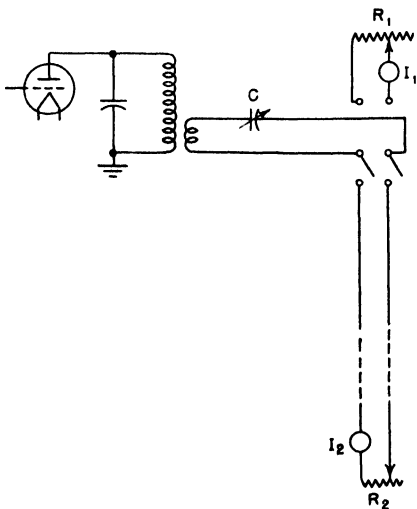



FIG. 81. Circuit for measuring impedance of transmission line.

the parallel two-conductor type, each conductor having a diameter of approximately ½ in. The spacing of the conductors is normally 12 to 15 in.

The curves of Fig. 80 show the characteristic impedance values with respect to spacing and conductor size of both the parallel-conductor-line and the concentric-tube type.

70. Transmission-line Calculations. There are diverse methods of measuring the characteristic impedance of a transmission line. A simple but effective method is illustrated in Fig. 81. With the setup shown and the switch thrown to the line position, a trial value of resistance R_2 is inserted. C is adjusted for maximum I_2. Then with switch thrown in the opposite position and R_1 set to equal R_2, C is adjusted

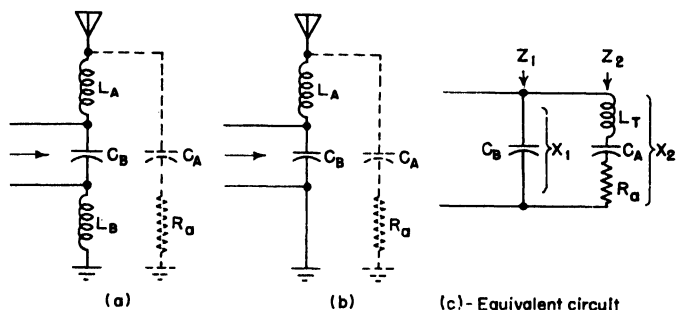



Fig. 82. Terminations for transmission lines.

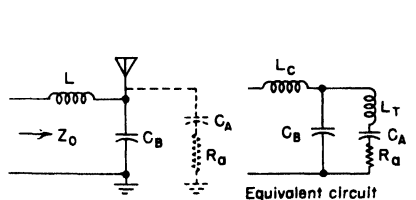



Fig. 83. Termination for transmission line with one side grounded.

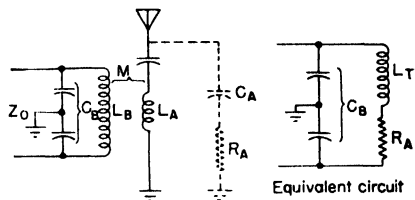



Fig. 84. Balanced transmission-line termination.

for maximum I_1. By trial, a combination may be found where there is a maximum value of I_1 and I_2 for the same setting of C with R_1 equal to R_2. This value of R_1 is the characteristic or surge impedance of the line.

When r-f power is transmitted over a transmission line to an antenna load, the line termination may be adjusted to afford a condition where there are no wave reflections by making the effective resistance of the termination equal to the characteristic impedance of the line. Several circuits used for terminating transmission lines are shown in Figs. 82 to 84 together with their equivalent circuits.

A formula for calculating the value of C_B for an effective resistance value Z_0 equal to the characteristic impedance of a two-conductor transmission line balanced to ground as shown in Fig. 82 as well as for a transmission line having one conductor grounded is as follows:

Let Z_0 = effective resistance of transmission-line termination
R_a = antenna resistance consisting of radiation resistance plus equivalent loss resistance
L_T = combined inductance-balance coils plus equivalent antenna inductance
C_A = equivalent antenna capacitance
C_B = line-termination capacitance
X_1 = reactance of C_B

X_2 = reactance of $X_L - X_{CA}$
Z_1 = impedance branch 1 = $-jX_1$
Z_2 = impedance branch 2 = $R_a + jX_2$

$$Z_0 = \frac{R_a X_1^2 - j(X_2^2 - X_1 X_2 + R_a^2)}{R_a^2 + (X_2 - X_1)^2}$$

$$C_B = \sqrt{\frac{Z_0 - R_a}{4\pi^2 f^2 Z_0^2 R_a}}$$

where $Z_0 > R_a$

$$X_1 = \frac{Z_0 R_a}{\pm \sqrt{R_a(Z_0 - R_a)}}$$

$$= \frac{1}{2\pi f C_B}$$

where, in Fig. 82, C_B is dependent only on values of Z_0 and R_a, where Z_0 is equivalent to a pure a-c resistance with the antenna circuit adjusted for resonance. *Unless Z_0 exceeds the value of R_a, an effective resistance equivalent to the characteristic impedance of the line cannot be secured.*

When low-impedance lines are used, such as the concentric-tube type, the termination shown in Fig. 83 is useful, since it affords a condition where correct termination may occur in the form of an effective resistance even though R_a equals or exceeds Z_0.

In Fig. 84 is shown a transmission-line termination in the form of a tank circuit. The tank capacitor C_B across the line is selected so as to provide a suitable kva ratio of the tank circuit with respect to the kilowatts transferred to the antenna circuit; this kva/kw ratio is normally about 10 and should never be less than 2.

$$X_1 = \frac{Z_0 R_e}{\pm \sqrt{R_e(Z_0 - R_e)}}$$

$$X_1^2 = \frac{Z_0^2 R_e}{Z_0 - R_e}$$

from which

$$R_e = \frac{X_1^2 Z_0}{Z_0 + X_1^2}$$

where R_e is the effective value of resistance reflected into the tank circuit from the antenna circuit.

The value of R_e can be calculated from

$$R_e = \frac{\omega^2 M^2 R_a}{R_a^2 + X_a^2}$$

where the inherent resistance of the tank circuit is negligible, M = mutual inductance between L_A and L_B, X_a = reactance of antenna circuit, and R_a = resistance of antenna circuit.

For a condition of proper termination X_a approaches zero and may be neglected and

$$\frac{\omega^2 M^2}{R_a} = \frac{X_1^2 Z_0}{Z_0^2 + X_1^2}$$

$$M = \sqrt{\frac{X_1^2 Z_0 R_a}{\omega^2 (Z_0^2 + X_1^2)}}$$

In Fig. 85 are shown values of M required for line impedances of 400, 500, and 600 ohms and a line-termination capacitor of between 0.001 and 0.004 μf. The

transmitter frequency was assumed as 670 kc and the antenna resistance as 30, 70, and 140 ohms. In the design of a tank-circuit termination for a given line the value of C_B across the line is selected so as to provide the proper kilovolt-amperes in the tank circuit with respect to the power transferred to the antenna circuit.

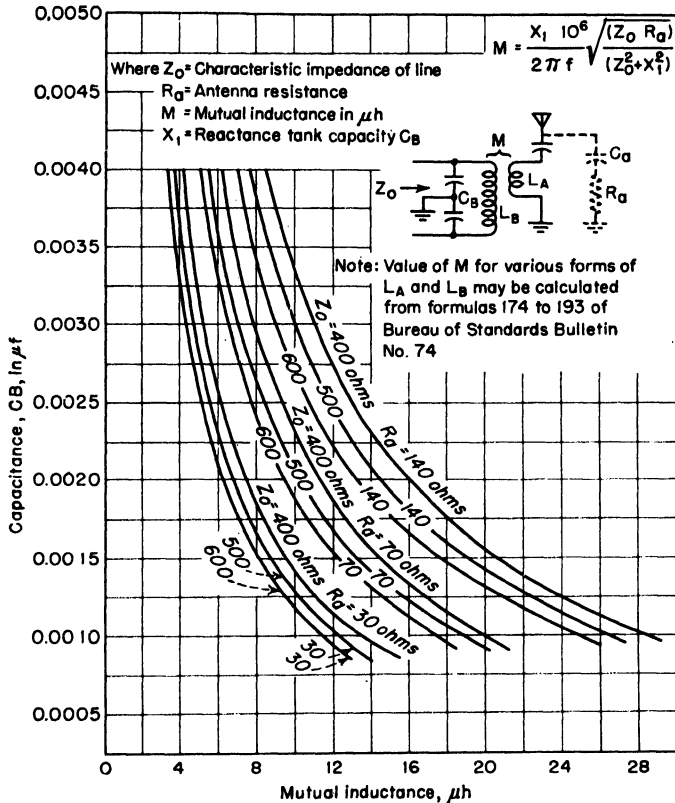



FIG. 85. Values of M required for proper termination.

71. Termination Adjustments. The usual procedure in adjusting a transmission-line termination for a condition of no wave reflection on the line is as follows:

1. The number of coupling turns is calculated so as to give the proper value of M. With the tank circuit open, the antenna is tuned to exact resonance by means of an external oscillator loosely coupled to it at the fundamental frequency.
2. The tank circuit is now connected into the circuit and tuned to resonance. This is indicated by a condition where the current in the antenna circuit becomes a minimum.
3. The transmission line is then connected across the tank circuit without making any changes in previous adjustments.
4. Correct termination may be checked by measuring the transmission-line currents at the ends and $\lambda/4$ points along the line. When proper termination has been effected, the transmission-line currents will be identical at all points along the line.

72. Concentric Line Terminations. The extensive use of concentric lines of the low-impedance type has led to cases where the characteristic impedance of the line is *lower* than that of the antenna resistance. In general there are three cases to consider,

as follows: (1) when the antenna impedance contains a resistance component only, (2) when the antenna impedance contains a resistance component and a reactance component, either (*a*) capacitive or (*b*) inductive, and (3) when the antenna impedance contains resistive and reactive components, the latter being partially compensated by the insertion of an extra reactance of opposite sign. These three cases are considered in order.

Case 1. *Antenna Impedance Purely Resistive.* From Fig. 86 the concentric-line characteristic impedance Z_0 is terminated by a network consisting of C_B, L_C, and the antenna

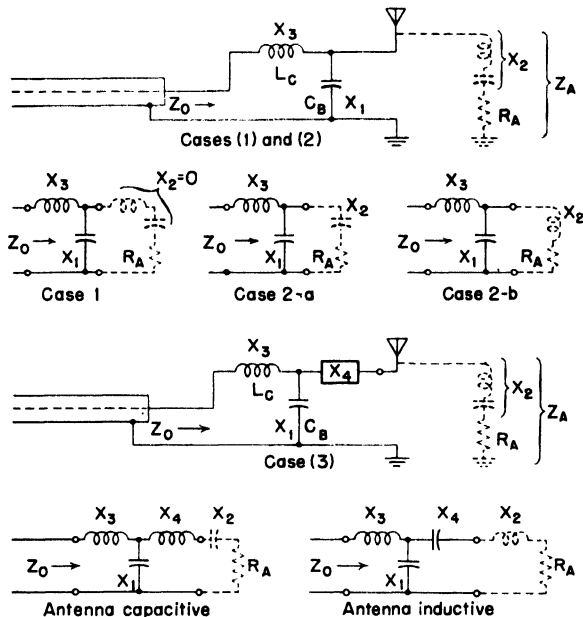



Fig. 86. Concentric-line terminations.

impedance Z_A. For case 1 the reactance of the antenna impedance is zero, and

$$Z_0 < Z_A = R_A$$

Then the complex impedance Z_L presented to the end of the transmission line is as follows:

$$Z_L = \frac{R_A[X_3X_1 - X_1(X_3 - X_1)] + j[X_3X_1^2 + R_A(X_3 - X_1)]}{R_A^2 + X_1^2}$$

where R_A, X_1, and X_3 are given in Fig. 86. For proper termination Z_0 must equal Z_L. X_1 becomes

$$X_1 = R_A \sqrt{\frac{Z_0}{R_A - Z_0}}$$

$$X_3 = \frac{R_A^2X_1}{X_1^2 + R_A^2}$$

Since $C_B = 1/(2\pi fX_1)$ and $L_C = X_3/(2\pi f)$, their values in microfarads and microhenrys are then readily calculable from f, the frequency of operation. Figure 87 gives various values of X_1 and X_3 in terms of values of Z_0 and R_A.

Case 2*a*. *Antenna Impedance with Capacitive Reactance.* Refer again to Fig. 86. It will be noted that the equivalent diagram for case 2*a* is the same as for case 1, except that

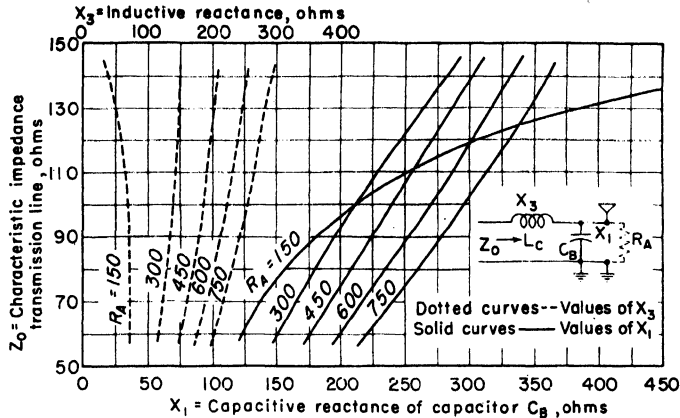



Fig. 87. Reactances required for proper line termination.

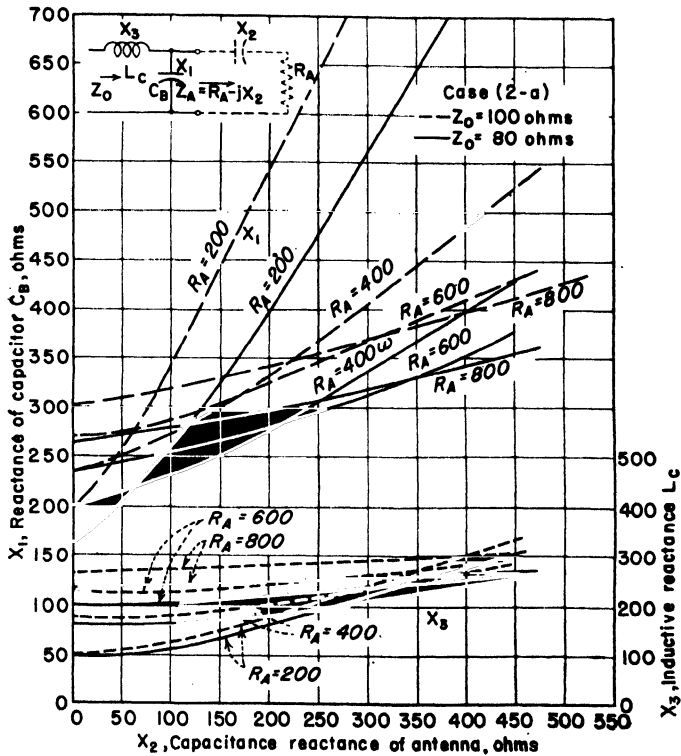



Fig. 88. Values of terminating reactances for case 2a.

the antenna impedance is now $Z_A = R_A - jX_2$. Then

$$Z_L = Z_0$$

$$Z_0 = \frac{R_A X_1^2}{R_A^2 + (X_2 + X_1)}$$

from which

$$X_1 = \frac{Z_0}{R_A - Z_0}\left[X_2 \pm \sqrt{\frac{R_A}{Z_0}(R_A^2 + X_2^2 - Z_0 R_A)}\right]$$

$$X_3 = \frac{X_1(R_A^2 + X_2 X_1 + X_2^2)}{R_A^2 + (X_1 + X_2)^2}$$

Hence, with Z_0, R_A, and X_2 given, X_1 and X_3 can be calculated. From the values of X_1 and X_3, L_C and C_B can be calculated, exactly as in case 1. Values of X_1 and X_3 for various values of R_A and values of X_2 for the cases where Z_0 is 80 and 100 ohms are given in Fig. 88.

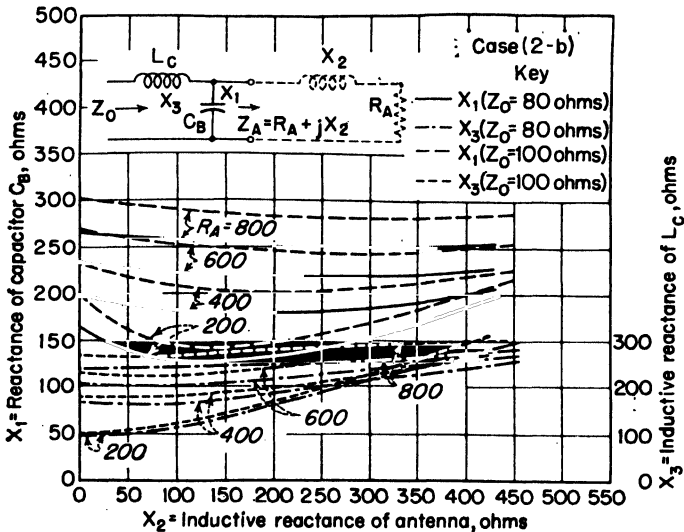



FIG. 89. Terminating reactances for case 2b.

Case 2b. Antenna Impedance Inductively Reactive. Case 2b is the same as case 1 except that $Z_A = R_A + jX_2$.

$$X_1 = \frac{Z_0}{R_A - Z_0}\left[-X_2 \pm \sqrt{\frac{R_A}{Z_0}(R_A^2 + X_2^2 - Z_0 R_A)}\right]$$

$$X_3 = \frac{X_1(R_A^2 + X_2^2 - X_2 X_1)}{R_A^2 + (X_2 - X_1)^2}$$

from which L_C and C_B are calculated. Figure 89 shows various values of X_1 and X_3 in terms of R_A and X_2, for Z_0 values of 80 and 100 ohms.

Case 3. Added Reactance to Antenna Impedance. When the transmission-line impedance "looks into" a complex antenna impedance, it is possible to simplify the adjustment of the circuit greatly by adding a reactance X_4 as shown in Fig. 86 for case 3. This reactive X_4 may be either inductive or capacitive, as shown. If the sum of X_3 and X_2 is inductive, then X_4 is made capacitive, and vice versa. The value of X_4 is such that the algebraic sum of X_1, X_2, and X_4 is equal to zero. Since X_4 is in series with the antenna impedance, it adds directly with the reactive part of the antenna impedance. The effect of the presence of X_4 can then be taken into account by applying the formulas of case 2a or 2b.

$$X_1 = X_3 = \sqrt{Z_0 R_A}$$

This occurs only, however, if X_4 is so chosen that

$$\pm X_4 - X_1 \mp X_2 = 0$$

The reactance X_4 must always have the opposite sign from X_2, as indicated by the plus-or-minus signs in the equation. When X_4 is so chosen, the reactance X_1 and X_3 may be obtained for various values of Z_0 and R_A by reference to Fig. 90. Note that these values apply regardless of whether R_A is larger than, equal to, or greater than Z_0.

Practical Procedure in Designing Matching Circuits. In making suitable adjustments on the impedance matching circuits to provide a correct termination for a given

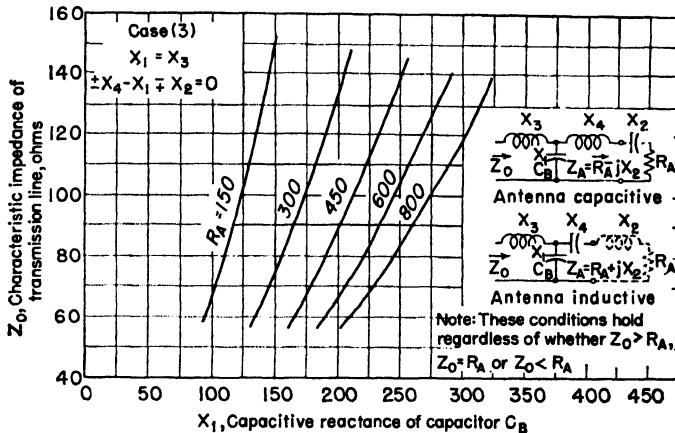



FIG. 90. Terminating reactances for case 3 when antenna reactance is compensated.

transmission-line characteristic impedance, under cases 2*a* and 2*b* above, where $R_A > Z_0$, the following procedure is recommended:

1. The transmission-line characteristic impedance should be calculated and the results checked by actual measurements if possible, either by means of an r-f impedance bridge or by the method described previously.
2. The antenna base resistance should be measured over a frequency band covering at least 100 kc each side of the operating frequency. A curve should then be constructed with values of antenna resistance as a function of frequency. A smooth curve drawn through the points of measurements will assist in checking their accuracy.
3. Together with antenna resistance measurements, the antenna reactance should be measured, either by means of an r-f impedance bridge or in a manner shown in Fig. 92 over a wide frequency range, and a curve constructed with antenna reactance as a function of frequency.
4. With the values of antenna resistance and reactance known, values of C_B and L_C may be calculated for case 2*a* or 2*b*, as may be required, and connected into the circuits as shown in Fig. 86.
5. With the transmission line connected, correct termination may be checked by measuring the transmission-line currents at the ends, if its length is equal to $\lambda/4$ or odd multiples thereof. For a very long line it is good practice to make these measurements at a number of points along the line. The existence of stationary waves of current or voltage of the fundamental frequency along the line is an indication of incorrect termination. In such a case slight adjustments may be necessary in L_C and L_B to correct for stray capacitance of leads and tuning equipment or slight errors in measurements. If an r-f impedance bridge is available, its measuring terminals may be connected across the input to the matching circuit in place of the transmission line and the termination circuit checked for an effective resistance equivalent to the characteristic impedance of the line without the line attached.

Case	Antenna current I_A	Current through capacitor C_B: $I_C\left(\frac{Z_A}{Z_A+Z_B}\right)=I_B$	Transmission line current $I_A\sqrt{\frac{R_A}{Z_0}}=I_C$	Voltage at base of antenna $E_A=I_AZ_A$
1	$\sqrt{\frac{W}{R_A}}$	$I_C\left[\frac{R_A^2+jR_AX_1}{R_A^2+X_1^2}\right]$	$\sqrt{\frac{W}{Z_0}}$	I_AR_A
2a	$\sqrt{\frac{W}{R_A}}$	$I_C\left[\frac{R_A^2+X_1X_2+X_2^2+jR_AX_1}{R_A^2+(X_2-X_1)^2}\right]$	$\sqrt{\frac{W}{Z_0}}$	$I_A(R_A-jX_2)$
2b	$\sqrt{\frac{W}{R_A}}$	$I_C\left[\frac{R_A^2-X_1X_2+X_2^2+jR_AX_1}{R_A^2+(X_2-X_1)^2}\right]$	$\sqrt{\frac{W}{Z_0}}$	$I_A(R_A+jX_2)$
3 Antenna capacitive	$\sqrt{\frac{W}{R_A}}$	$I_C\left[\frac{R_A+jX_1}{R_A}\right]$	$\sqrt{\frac{W}{Z_0}}$	$I_A(R_A-jX_2)$
3 Antenna inductive	$\sqrt{\frac{W}{R_A}}$	$I_C\left[\frac{R_A+jX_1}{R_A}\right]$	$\sqrt{\frac{W}{Z_0}}$	$I_A(R_A+jX_2)$
W is power in watts				

FIG. 91. Current and voltage relations in transmission-line terminating circuits.

Although case 3 requires the addition of another piece of apparatus in the form of an inductance or capacitance in the antenna lead, which may be rather expensive, the adjustment procedure is less difficult and is as follows:

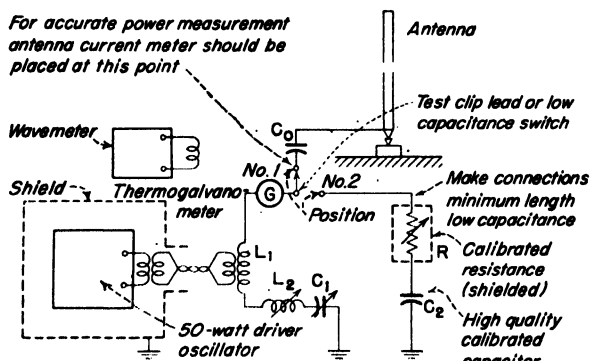



FIG. 92. Setup for measuring antenna characteristics.

1. With values of the line characteristic impedance, antenna resistance, and reactance obtained by measurement, the value of C_B is calculated, which gives the reactance X_1 necessary.

2. With L_C disconnected from C_B, X_4 (inductive or capacitive) is added in the antenna circuit in series with X_1. By means of X_4 the antenna circuit is tuned to resonance, as indicated by maximum current through a thermogalvanometer, when the antenna circuit is excited by means of an external oscillator loosely coupled to it.

3. A sufficient value of inductance L_C having a value X_3 equal to X_1 is then connected into the circuit as shown in Fig. 86.

4. The line is then checked for stationary waves, the absence of which indicates a condition of correct termination.

The mechanical properties of long concentric-tube transmission lines makes the measurement of current in the center conductor rather difficult. In some cases removable plugs are placed in the outside tube at various intervals along the line. These plugs, which, when inserted, make the outer tube airtight, permit connections from an antiresonant circuit across the line. Such an antiresonant circuit, when tuned to the fundamental frequency, presents a very high impedance to the line, when bridged across it, and therefore does not affect its characteristic impedance at the fundamental frequency. With about 10 watts flowing through the line, the galvanometer reading is an indication of the voltage at the points measured along the line.

Fig. 93. Matching circuits for balanced transmission line.

73. Method Used in Measuring Antenna Characteristics. Refer to Fig. 92.

C_0 (usually about 0.0005 μf) is selected to provide sufficient series capacitance-reactance to make the antenna capacitive over the frequency range measured. Then, with the antenna excited by the driver oscillator at the frequency indicated by the wavemeter and the switch at position 1, adjust C_1 and L_2 for resonance, as indicated by the maximum reading of G. R is then adjusted until the G reading is the same as before. Then R is the antenna resistance.

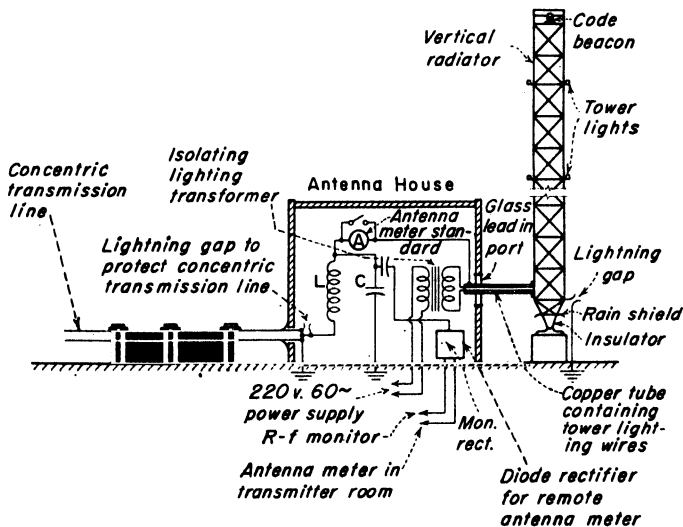
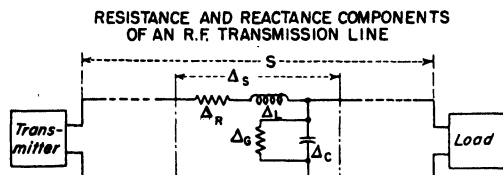
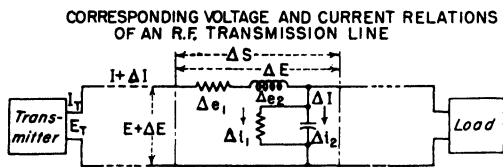



Fig. 94. Equipment in antenna house of modern broadcast transmitter.

For antenna reactance measurement, the circuit is first calibrated for stray capacitance in the shielded resistance box by resonating circuit (switch in position 2) first with box in the circuit and then entirely removed. Difference in reading of C_2 between the two conditions

equals capacitance of box. This value should be added to each reading of C_2, when circuits are resonated, which is done as above for resistance measurement. The antenna reactance X_2 is equal to the reactance of C_2 minus that of C_0. When reactance of C_0 is greater than that of C_2 the antenna reactance is positive.

When it is found desirable to apply the matching circuits described above (Fig. 86) to balanced lines (open wire or double concentric types), the value of X_3 derived by the particular formula for cases 1 and 2 is halved and placed on each side of the circuit (see Fig. 93), while the value of X_4 is halved and placed on each side of the circuit for case 3. Under these conditions the formulas given above apply to the respective cases mentioned. The systems become quite useful in matching a given balanced transmission line or r-f circuit into another having entirely different input impedance characteristics.

In the foregoing analysis of antenna matching circuits, they were considered as providing, for a given transmission line, a termination impedance equivalent to an ohmic resistance at the fundamental frequency. An analysis of the input impedance that such a line "looks into" at various harmonic frequencies discloses that it may assume an infinite number of different impedances containing resistance and positive or negative reactance components, the values of which depend upon the termination circuit constants as well as those of the antenna. The values of antenna resistance and reactance may vary widely with frequency. For harmonic frequencies, stationary waves of current and voltage will form on the transmission line as well as in the antenna circuit, unless suitable harmonic filtering is provided either within the vacuum-tube transmitter or at the input to the transmission line.

Relationship along line any variation voltage or current

$$\frac{\partial I}{\partial S} = EG + C\frac{\partial E}{\partial t} \qquad \frac{\partial E}{\partial S} = IR + L\frac{\partial I}{\partial t}$$

Assume sinusoidal variation of the current and voltage at any point along the line a distance S from transmitter

$$I = I_T \cosh S\sqrt{ZY} - E_T\sqrt{Y/Z}\sinh S\sqrt{ZY} \qquad E = E_T \cosh S\sqrt{ZY} - I_T\sqrt{Z/Y}\sinh S\sqrt{ZY}$$

When the line is aperiodic surge impedance $Z_0 = \sqrt{\frac{R+j\omega L}{G+j\omega C}}$

Components contributing to line loss:

Δe_1 = Voltage consumed in phase with current

Δi_1 = Current consumed in phase with voltage

Voltage E_S at any point S distance from transmitter end

$$E_S = E_T e^{-s(\alpha+j\beta)} = E_0 e^{-S\sqrt{(R+j\omega L)(G+j\omega C)}}$$

where α = attenuation constant for lines of small leakage conductance $\alpha = \doteq \left(\frac{R}{2} + \frac{GL}{2C}\right)\sqrt{C/L}$

For lines having negligable leakage $\alpha \doteq \frac{R}{2}\sqrt{\frac{L}{C}} = \frac{R}{2Z_0}$

$A \doteq 4.343\frac{R}{Z_0}$ decibels per unit length

FIG. 95. Losses in transmission lines.

The effectiveness of a given filter for various harmonics depends upon its position in the line with respect to the positions of current and voltage antinodes of the harmonic frequencies along the line.

74. Loss in R-f Transmission Lines. From Fig. 95, it is evident that the most prominent factors contributing to power loss in open-wire transmission lines are as follows:

1. Power loss due to conductor thermal resistance for two-wire line

$$R_T = \frac{0.1262}{d}\sqrt{\rho\mu f} \qquad \text{ohms per cm length}$$

where $S >> d$ (see Fig. 80)

ρ = resistivity of conductors, in ohms per cm cube
μ = permeability of conductors
f = frequency, Mc
d = diameter of conductor, cm

2. Power radiated from balanced and unbalanced line currents
3. Power component of mutual inductance due to secondary currents induced in nearby conductors
4. Power loss due to leakage or conductance of the insulating medium or

$$G = \frac{\gamma A}{l} \qquad \text{mhos per cm length}$$

5. Power loss due to dielectric hysteresis

NOTE: The resistivity of copper is approximately 1.724×10^{-6} ohm per cm cube at 20°C. The exact value of a particular grade of copper depends upon the amounts of other elements it may contain.

For coaxial lines the major factors contributing to power loss in transmission lines are as follows:

1. Power loss due to conductor thermal resistance

$$R_T = 0.0631\sqrt{\rho\mu f}\left(\frac{1}{d_1} + \frac{1}{d_2}\right) \qquad \text{ohms per cm of line}$$

where d_1 = outside diameter of inner conductor, cm
d_2 = inside diameter of outer conductor, cm
ρ = resistivity of conductors, in ohms per cm cube
μ = permeability of conductors
f = frequency, Mc

2. Power loss due to leakage conductance of insulating medium
3. Power loss due to dielectric hysteresis

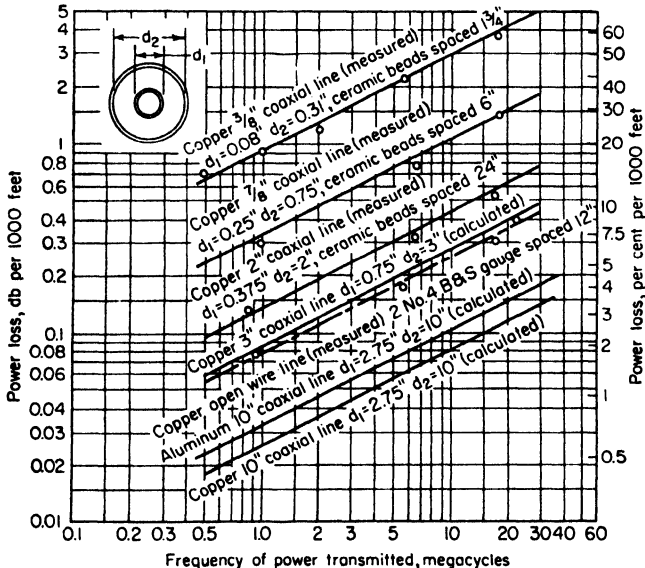



FIG. 96. Power loss in lines operating at 20°C with negligible reflection.

By reference to the loss curves shown in Fig. 96, it is evident that the greater part of the power loss in both the open-wire and coaxial types when operating with negligible reflection is due to the conductor thermal resistance. Owing to the low-loss insulation materials now available, the losses due to dielectric leakage and hysteresis can be reduced to a negligible quantity especially at standard broadcast frequencies.

The curves of Fig. 96 are the results of extensive r-f measurements which confirmed mathematical formulas given for calculation of losses in open-wire and coaxial lines of various standard sizes. Worthy of mention is the rather high efficiency of the open-wire line consisting of two No. 4 B&S gage solid copper conductors spaced 12 in. center to center. The losses in this line are less than those in a 3-in.-diameter copper coaxial line. The resistivity ρ of aluminum being greater than copper, the losses in an aluminum coaxial line are somewhat greater than those in a copper line of the same dimensions.

FIELD STRENGTHS REQUIRED AND MEASURED

75. Broadcast-station Signal Coverage. The reception of satisfactory signals from a given broadcasting station by a particular listener at a given point depends upon the following: (1) the intensity of the signal radiated from the antenna system of the station as influenced by the radiated r-f carrier power, antenna directivity, and percentage of modulation on the carrier; (2) distance between the transmitting antenna and the point of reception and the attenuation characteristics of the intervening space or terrain; (3) intensity of objectionable interference at the receiving point; (4) fading as produced by the rays of direct and indirect signals; (5) the quality of the receiver and its ability to discriminate against local noise or interference and against adjacent channel interference and to convert the receiver r-f signals into sound without appreciable distortion or inherent receiver noise. The surrounding area about a given station wherein satisfactory program signals can be received determines the service area of the station.

The service area of a station of the 550- to 1,600-kc band consists essentially of two distinct regions. The region in close proximity to the station is served by the direct ray or ground wave called the *primary coverage area*, while the region at some distance from the station and served by virtue of indirect ray or sky-wave reflections is called the *secondary coverage area*. During daylight hours on frequencies between 550 and 1,600 kc, a broadcast listener is concerned with the primary coverage area signals of nearby stations since there is very little sky-wave energy reflected during this period under normal conditions. The daylight service area of such a station therefore consists almost entirely of that region served by the direct ray.

During the hours of twilight and darkness, the secondary coverage area of stations in the standard broadcast band becomes apparent. The secondary coverage area of a particular station begins at a considerable distance from a given station and is served by the predominant sky wave. The primary and secondary coverage areas are separated by a region known as the *fading area*. In this area the signal intensities of the direct and indirect rays approach an equality with a result that violent fluctuations in signal intensities are apparent. The fading areas are dependent upon a number of factors, such as frequency, antenna, radiation characteristics, conductivity of intervening terrain, and time of day and season, and are independent of the transmitter powers of the stations. The fading area is normally in the form of a band about the station normally contained within radii of between 20 and several hundred miles, depending upon the factors mentioned. The fading band may be as much as 50 miles in width.

Considering a standard broadcasting station radiating equally in all directions over surrounding terrain, and assuming equal ground attenuation, the service area would consist of a primary coverage area near the station and served by a steady ground-wave signal. Outside of this would exist the fading area consisting of a ring about the primary area. Beyond the fading ring the secondary coverage area will exist. Inasmuch as broadcast reception is rather uncertain in the fading region and in the

secondary coverage area, the real value of a given station is dependent normally upon its primary coverage area.

The primary service area of a particular station can be most accurately determined by means of a field-intensity survey. A survey[1] of this kind is accomplished through the use of mobile field-intensity measuring equipment. This consists essentially of a field-intensity meter or carefully shielded receiver equipped with an indicating meter at its output terminals to read carrier-signal intensity as induced in the loop antenna. The field-intensity meter and the loop antenna are carefully calibrated in their position in the measuring car to give accurate readings in microvolts per meter over a wide range of carrier-signal intensity.

76. Field-intensity Measurements. The procedure of making a field-intensity survey of a broadcasting station consists usually of making frequent measurements at satisfactory positions (in free space) along radials progressing to and from the station. Eight or more radials at equal angular spacing are generally made about a point established on the field survey map by the station antenna system and extending to a signal intensity of 500 mv or beyond. Each radial is then plotted on loglog paper and a smooth curve drawn through these points to show directly the signal intensity along one ordinate, with distance along the other. Later the values required are transferred to a map in the form of signal contour lines representing positions about the station where field intensities of 100, 50, 10, 2, and 0.5 mv per m exist. The contour map for reference purposes also contains information such as (1) station call letters, (2) frequency, (3) antenna power and its directivity and other characteristics, (4) scale of map, and (5) date.

Since fading occurs after sunset these measurements are an indication of satisfactory daytime coverage only from the particular station. As recommended by reports of the IRE,[2] FCC,[3] and the National Association of Broadcasters, values of standard broadcast field intensity considered necessary for reliable broadcast service are given for three areas as follows: (1) a business city area where a field intensity of from 10 to 50 mv per m is required to override high interfering electrical noise and overshadowing effects of large buildings, (2) a residential district of a city where a field intensity of 2 to 10 mv per m is required, (3) a rural area where 0.1 to 0.5 mv per m signal intensity is sufficient. In addition it is stated that for fair service a signal intensity of one-half the above values is needed and for poor service one-fourth of these values. These figures are based upon the average signal intensity necessary to override the noise levels of these districts. In large cities, where large, tall buildings are numerous, a free-space field intensity of as much as 50 mv per m over the city may be necessary to provide a signal intensity at a particular receiving antenna between buildings.

Since the primary service area of a standard broadcasting station includes nighttime reception as well as daytime, fading measurements are necessarily a part of the field-intensity survey in determination of this area. Fading measurements are made with the same field-intensity measuring equipment used for the survey except that the field-intensity meter is equipped with a recording milliammeter (usually of 0 to 5 ma range) attached to the output of the field-intensity meter. A d-c amplifier sometimes is necessary to secure sufficient signal level to actuate the recording meter from the field-intensity measuring set. The equipment is set up for periods of time at a given distance and location from the station, and fluctuations in carrier-signal intensity are noted on the continuously moving recording chart. Amplitude fluctuations as recorded on the chart indicate the amount of fading. Fading measurements for considerable periods of time and over a wide area are necessary to determine the fading region about a given station and to evaluate the secondary coverage area about the station, particularly those designated as class I stations.

[1] Kirby, S. S., and K. A. Norton, Field Intensity Measurements, *Bur. Standards J. Research*, April, 1932.

[2] Report of Committee on Radio Propagation Data, *Proc. IRE*, **21** (10), October, 1933.

[3] Fifth annual report to the Congress of the United States, by Federal Communications Commission, gives tabulated values of field strength.

Field-intensity measurements for v-h-f broadcasting stations are made in a manner similar to that for standard broadcasting with the exception that a different type of field-intensity measuring equipment is required and that the procedure for making fading measurements is not generally required.

77. Calculations of Station Coverage. A mathematical investigation of the attenuation of radio waves propagating over plane earth has led to expressions which follow very nearly the characteristics of waves as indicated by actual measurements. A simplified form of this expression requires the following information for a solution: (1) frequency, (2) distance from the station, (3) conductivity of the soil in electromagnetic units, and (4) inductivity of the soil in electrostatic units. Since the inductivity can be generally assumed to be 14 to 15 esu, then, with a measured value of conductivity σ, the field intensity at a given distance from a station may be calculated. With further assumptions concerning the irregularities in general characteristics of the terrain about the station, it is possible to calculate the contours. The value of σ (the soil conductivity) is usually secured from a measured radial or taken from available field-intensity measurements of some other station in the vicinity. The FCC has published charts showing soil conductivity over the United States. Provided measured values are not available, these may be used.

For convenience the chart shown in Fig. 97 is given. It may be used to calculate signal attenuation of standard broadcast frequencies. The attenuation curves shown are derived from a simplified form of Sommerfeld's attenuation formula.[1] With a single set of Sommerfeld curves to cover all the standard broadcast frequencies and soil conductivities, the conductivity of a given soil can be rather easily computed from the attenuation of a particular signal. This is accomplished by first converting a given radial to an inverse field strength of 1,000 mv at 1 mile and then determining the frequency of the ground-wave curve with which it coincides, *i.e.*, the conversion frequency. The conductivity is secured from the soil-constant curve passing through the intersection of the operating and conversion frequencies on the conversion chart in the upper right corner of the figure. The conversion chart has been prepared from the following relationships:

$$f_1 = f\sqrt{\frac{\delta}{\delta_1}}$$

where f = operating frequency
δ = standard conductivity of chart (100×10^{-15} emu)
δ_1 = actual soil conductivity
f_1 = conversion frequency

For example, assume a station operating on 660 kc where the field strength radial, as plotted from measurements, follows the 1,500-kc curve. Then, from the soil conversion chart, the conductivity is very nearly 20×10^{-15} emu. On the other hand, if the soil conductivity is known, the signal attenuation can be determined from the conversion chart and the attenuation curves of various frequencies. Since these curves are based on a field strength of 1,000 mv at 1 mile, the actual signal at a given distance from a station is, of course, derived from the ratio of the actual signal intensity in millivolts at 1 mile from the particular station divided by 1,000. At considerable distances from the transmitter these curves are subject to corrections for the effects of curvature of the earth.

The curves in Fig. 98 refer to sky-wave intensities under various conditions of

[1] Sommerfeld, Arnold, Ausbreitung der Wellen in der drahtlosen Telegraphie Einfluss der Bodenbeschaffenheit, und gerichtete und ungerichtete Wellenzüge, *Jahrb. drahtlosen Tele. Tele.*, **4**, December, 1910. Rolf, Numerical Discussion of Sommerfeld's Attenuation Formula, *Proc. IRE*, **18** (3), March, 1930. Eckersley, P. P., The Calculation of the Service Area of Broadcast Stations, *Proc. IRE*, **18** (7), July, 1930. Eckersley, T. L., Direct Ray Broadcast Transmission, *Proc. IRE*, **20** (10), October, 1932. Norton, K. A., Propagation of Radio Waves over a Plane Earth, *Nature*, June 8, 1935; Propagation of Radio Waves over the Surface of the Earth and in the Upper Atmosphere, Part I, *Proc. IRE*, **24**, October, 1936; Part II, *Proc. IRE*, **25**, September, 1937. Fitch, W. A., The Sommerfeld Formula, *Electronics*, **9** (9), September, 1936. For FCC curves of standard broadcasting-station propagation, refer to curves of FCC rules, Part 3, Secs. 3.184 and 3.186, U.S. Government Printing Office, Washington.

propagation. These are most useful in the determination of the fading regions about a particular station and are plotted to give intensities of reflected sky-wave intensity for different antenna electrical heights based on a signal intensity of 1,000 mv along the ground at 1 mile from a given antenna. In this case the electrical height of the

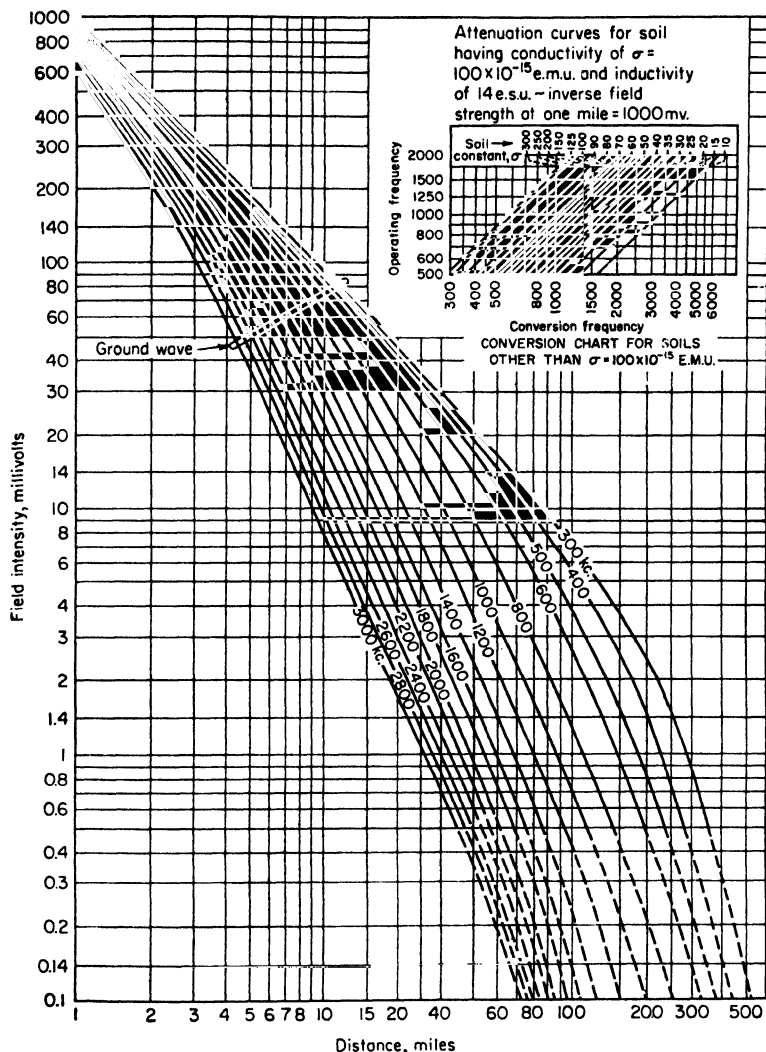



FIG. 97. Ground-wave field-intensity curves.

antenna in degrees equals $3.85 \times 10^{-4}Hf$, where H is the physical height of the antenna in feet and f is the operating frequency in kilocycles. This is based on a velocity of propagation equivalent to 0.95 that of light.

Inasmuch as the attenuation curves of ground-wave intensity (Fig. 97) are also based on 1,000 mv at 1 mile, then the particular distance from a given antenna where the sky-wave intensity, shown on curves of Fig. 98, equals the ground-wave signal

intensity of Fig. 97 is the distance from the antenna where one would expect to observe greatest fading or is an estimate of the center of the fading band. Owing to the height of the Heaviside layer being other than 100 km and reflection being less than unity, on which these curves are based, calculated distances given by these curves

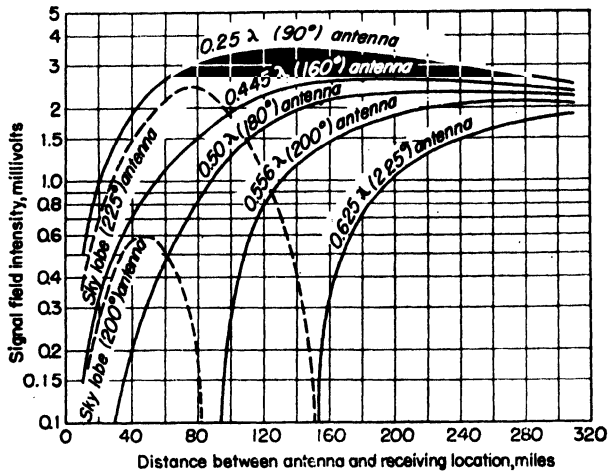



FIG. 98. Sky-wave propagation curves for antennas of various heights. Layer height = 100 km; reflection coefficient = 1.

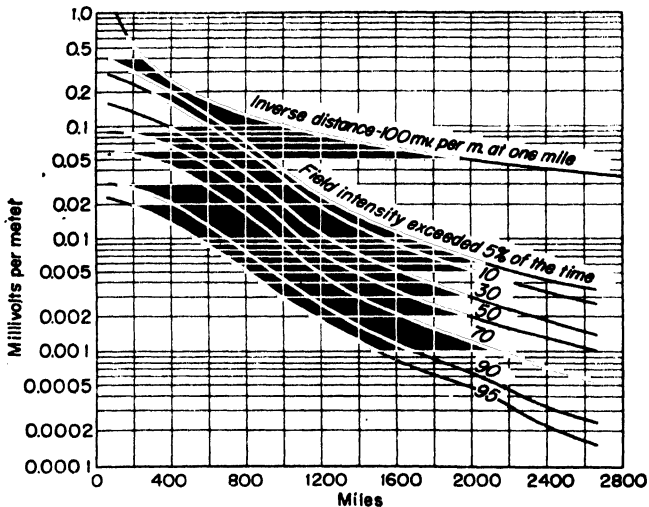



FIG. 99. Average sky-wave field intensity, 640 to 1,190 kc.

are approximate. Measurements are required for more exact determination of the fading region.

The service rendered by a standard broadcast station depends also on interference caused by other stations on the same and nearby channels. This interference is greatly increased at night because signals from undesired distant stations are reflected by the Heaviside layer and may be received with varying intensities within the service

area of a desired station. Following extensive survey work covering nighttime signal propagation over the period February to May, 1936, the FCC issued a report[1] wherein a great amount of information concerning sky-wave propagation is given. In Fig. 99 are illustrated curves representing the average sky-wave field intensity (second hour after sunset) at the recording station. An interfering or undesired signal existing for 10 per cent of the time has been standardized as an interfering signal. Thus, with the antenna sky-radiation characteristics of a given station known,[2] it becomes possible to estimate the amount of undesirable interference it is liable to cause to another distant station. In the determination of interference problems, the FCC has issued under Sec. 3.182W of its rules Table 4 to be used for determining the minimum ratio of the field intensity of a desired to an undesired signal for interference-free service.

Table 4. Interference Ratios

Frequency separation of desired to undesired signals, kc	Desired ground wave to		Desired 50% sky wave to undesired 10% sky wave
	Undesired ground wave	Undesired 10% sky wave	
0	20:1	20:1	20:1
10	1:1	1:5	*
20	1:30		

* The secondary service area of a class I station is not protected from adjacent channel interference. However, if it is desired to make a determination of the area in which adjacent channel ground-wave interference (10 kc removed) to sky-wave service exists, it may be considered as the area where the ratio of the desired 50% sky wave of the class I station to the undesired ground wave of a station 10 kc removed is 1:4.

When the desired ground-wave signal is interfered with by two or more sky-wave signals on the same frequency, the rss value of the latter is used.

From the above, it is apparent that in many cases stations operating 10 and 20 kc apart may be operated with antenna systems side by side or otherwise in proximity without any indications of interference if the interference is defined only in terms of permissible ratios listed above. As a practical matter, serious interference problems may arise when two or more stations with the same general service area are operated on channels 10, 20, and 30 kc apart.

The FCC has classified standard broadcast stations with respect to protected service contours and permissible interference signals in accordance with Table 5 under Sec. 3.182V of the FCC rules.

78. V-h-f broadcast-station coverage concerns the stations licensed by the FCC primarily for transmission in the v-h-f broadcast band for reception by the general public. The v-h-f broadcast band extends from 88 to 108 Mc. Stations in this band must use f.m. The assigned operating frequency or "center frequency" is that of the r-f carrier without modulation. It must be maintained within 2,000 cps of the assigned center frequency assigned. Channels for v-h-f broadcast stations begin at 88.1 Mc and continue in successive steps of 200 kc to and including the assigned frequency of 107.9 Mc.

According to the FCC rules, v-h-f broadcast stations shall be licensed on the basis of an area in square miles within the service area. The contour bounding the service area and the radii of same are determined in accordance with the FCC standards. On this basis, a v-h-f broadcast station has a single service, that corresponding to the primary service of a standard broadcast station. Secondary (sky-wave or intermittent) service is not recognized in v-h-f broadcast coverage.

[1] FCC Report 18108, September, 1936.
[2] FCC rules, Sec. 3.190, Fig. 1.

Table 5. Protected Service Contours and Permissible Interference Signals for Broadcast Stations

Class of station	Class of channel used	Permissible power, kw	Signal-intensity contour of area protected from objectionable interference,[a] μv/m		Permissible interfering signal on same channel,[b] μv/m	
			Day[c]	Night	Day[c]	Night[d]
I-A	Clear	50	SC 100 AC 500	Not duplicated	5	Not duplicated
I-B	Clear	10–50	SC 100 AC 500	500 (50% sky wave)	5	25
II	Clear	0.25–50	500	2,500 (ground wave)	25	125
III-A	Regional	1–5	500	2,500 (ground wave)	25	125
III-B	Regional	0.5–1 night and 5 day	500	4,000 (ground wave)	25	200
IV	Local	0.1–0.25	500	Not prescribed[e]	25	Not prescribed[e]

[a] When it is shown that primary service is rendered by any of the above classes of stations, beyond the normally protected contour, and when primary service to approximately 90% of the population (population served with adequate signal) of the area between the normally protected contour and the contour which such station actually serves is not supplied by any other station or stations, the contour to which protection may be afforded in such cases will be determined from the individual merits of the case under consideration. When a station is already limited by interference from other stations to a contour of higher value than that normally protected for its class, this contour shall be the established standard for such station with respect to interference from all other stations.

[b] For adjacent channels see appropriate paragraph of the FCC rules.

[c] Ground wave.

[d] Sky-wave field intensity for 10% or more of the time.

[e] These values are with respect to interference from all stations except class I-B, which stations may cause interference to a field-intensity contour of higher value. However, it is recommended that class II stations be so located that the interference received from class I-B stations will not exceed these values. If the class II stations are limited by class I-B stations to higher values, then such values shall be the established standard with respect to protection from all other stations.

V-h-f broadcast transmitters are normally located as near to the center of the proposed service area as possible. A high elevation of the transmitting antenna is necessary to reduce the shadowing effects upon propagation due to hills, buildings, and other obstructions in the city area. The position of the transmitter site is also dependent upon the purpose of the station, *i.e.*, whether it is intended to serve a small city, a metropolitan area, or a large region. A suitable transmitter site may be made available by the use of a directive antenna. Where a directive antenna is used, a centrally located station site may not be a desirable one. As one may understand by studying the chart in Fig. 100, the transmitter antenna height above the average elevation of the service area is a consideration of greatest importance to secure optimum coverage.

Table 6. Allowable Signal Ratios

Channel Separation	Ratio of Desired to Undesired Signals
Same channel	10:1
Adjacent channel (200 kc removed)	2:1
400 kc removed	1:10
600 kc removed	1:100
800 kc removed and above	No restriction

In the consideration of objectionable interference from other stations on the same and adjacent channels, the FCC[1] rules require that the proposed station shall not have interference to such an extent that its service may be reduced to an unsatisfactory amount. For this reason, objectionable interference is considered to exist when the interfering signal exceeds that given by Table 6. In this table the desired signal is

[1] FCC Rules 3.313b.

the median field intensity, and the undesired signal is the tropospheric signal intensity exceeded for 1 per cent of the time.

The FCC[1] rules concerning frequency-carrier separation of v-h-f stations further state:

Stations normally will not be authorized to operate in the same city or in nearby cities with a frequency separation of less than 800 kc: *Provided*, That stations may be authorized

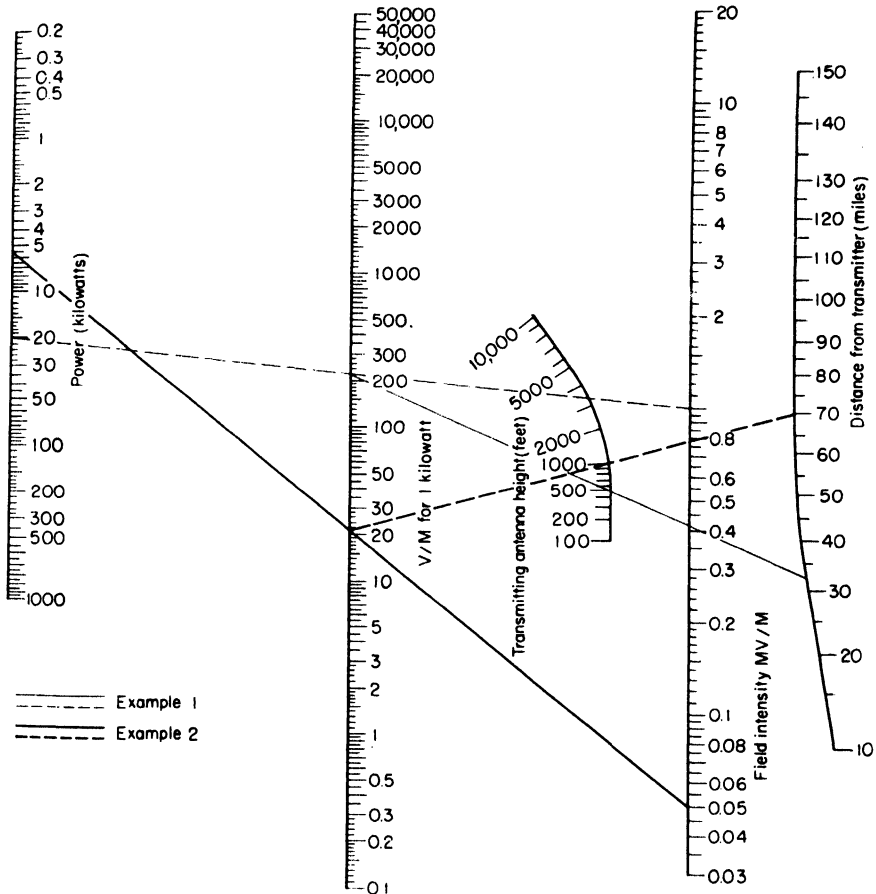



Fig. 100. Signal range for v-h-f broadcast stations.

to operate in nearby cities with a frequency separation of not less than 400 kc where necessary in order to provide an equitable and efficient distribution of facilities: *And provided further*, That class B stations will not be authorized in the same metropolitan district with a frequency separation of less than 800 kc. In the assignment of FM broadcast facilities the Commission will endeavor to provide the optimum use of the channels in the band, and accordingly may assign a channel different from that requested in an application.

Objectionable interference is not considered to exist when the channel separation is 400 kc or greater. Therefore f-m broadcast stations in the same city or area may be assigned channels 400 kc apart.

[1] FCC Rules 3.313c.

The service area of f-m broadcasting stations is considered to be only that served by the ground wave. The extent of the service is determined by the point at which the ground wave is no longer of sufficient intensity to provide satisfactory service. The standard of field intensity necessary for satisfactory service is given as follows:

Table 7. Service

Area	Median Field Intensity, Mv/m
City areas near factories, car lines, or busy streets	1
Rural areas away from highways	0.050

The FCC rules[1] state that

. . . a median field intensity of 3,000 to 5,000 μv per m should be placed over the principal city to be served and a median field intensity of 1,000 μv per m should be placed over the business district of cities of 10,000 or greater within the metropolitan district served.

To determine the approximate distance to a particular contour of a v-h-f f-m station the chart in Fig. 100 may be used. The results obtainable are based on a signal intensity at a receiving antenna with an elevation of 30 ft. The distance to the 50 μv per m contour about a given station is dependent upon values of the transmitting antenna height, the antenna power, and the antenna gain. This chart is prepared for a carrier frequency of 98 Mc in the center of the f-m band and may be used for all f-m broadcast channels in the 88- to 108-Mc band with sufficient accuracy since little change results over this frequency range.

Examples: Consider (1) a northeastern metropolitan station and (2) a transmitter site 1,000 ft high with a 50-μv signal required at a distance of 70 miles from the transmitting antenna.

1. The radiated power of the northeastern station is assumed to be 20 kw at an antenna height of 500 ft. The problem is to predict the range in miles to the 1-mv contour. In Fig. 100, the dashed line between 20 kw and the 1 mv per m field intensity is drawn. This line intersects the microvolts for 1 kw line at 224 μv. A solid line then drawn between this point of intersection and the 500-ft transmitting-antenna-height point shows a distance of 32 miles.

2. Assume that there is required a 50-μv signal 70 miles away. Draw, between the 70-mile point and the point indicating 1,000 ft, a line which indicates that a radiated power of 1 kw from a station using an antenna 1,000 ft high would provide 21.5 μv at a distance of 70 miles. A solid line drawn between this point, the 0.050-mv point, and the power scale indicates that a transmitter power output of 5.5 kw is required to give a field intensity of 50 μv (0.050 mv) at a distance of 70 miles from the antenna radiating at an elevation of 1,000 ft above average ground at the transmitting station.

[1] FCC Rules 3.311b.

CHAPTER 22

TELEVISION

By Raymond F. Guy[1]

1. Definition. Television is the electrical transmission of transient visual images. Although the process is fundamentally simple an actual television system is one of great complexity.

2. Elements. A television camera tube employing an electron scanning beam reads off variations of voltage amplitude, corresponding to brightness, from a photosensitive surface upon which a picture image is focused. The image is recreated in a receiver picture tube when a corresponding synchronously modulated and deflected electron beam impinges upon a phosphorescent surface. Television systems normally contain facilities for synchronous transmission and reproduction of visual images and sound, but in some industrial applications the sound system may be omitted.

Figure 1 shows the elements of a typical system. The sound system consists of an f-m transmitter and receiver in which, under the FCC standards, a maximum frequency deviation of 25 kc corresponds to 100 per cent modulation. The sound system is operated in parallel with and separate from the picture system, except that common sound and picture antennas may be employed at both the transmitter and receiver and a common r-f amplifier, first detector, i-f amplifier, and second detector may be used in the receiver.

The picture transmitter includes the camera and synchronization circuits which create the video signals, video amplifiers, an r-f carrier source and r-f amplifiers, the modulator, a vestigial-side-band filter for suppressing part of the lower-frequency side bands from the transmitted signals, a sound and picture-signal diplexer for channeling both sound and picture signals to a common antenna, and the antenna.

The type of receiver most commonly used processes the sound and picture signals together up to the output of the second detector through antenna-to-receiver coupling circuits, a first detector, an i-f amplifier, and a second detector. Thereafter, there is a video amplifier, synchronizing-signal separator circuits, and a picture tube for the picture signals and a separate 4.5-Mc i-f amplifier, f-m sound demodulator, audio amplifier, and loud-speaker for the sound signals. The sound carrier appears at the output of the second detector as an f-m subcarrier of 4.5 Mc, which is the separation between the transmitted and received picture and sound carriers.

The receiver described and shown in Fig. 1 is referred to as the "intercarrier" type and is now more commonly used than older types which employed separate sound and picture i-f amplifiers and second detectors following the first detector.

SCANNING AND IMAGE ANALYSIS

3. Linear Scanning. The method of analyzing and synthesizing visual images employed in modern television systems is known as *linear scanning*. As applied to the transmission of images, linear scanning involves the exploration of the image by an elemental spot of small area, known as the *scanning agent*, which traverses the area of the image in a series of horizontal lines, moving over every point in the image at constant speed and discovering the degree of brightness at each point in

[1] Director of Radio Frequency and Allocations Engineering, National Broadcasting Company, Inc.

NOTE: This chapter draws freely on material written for the fourth edition of this book by Donald G. Fink.

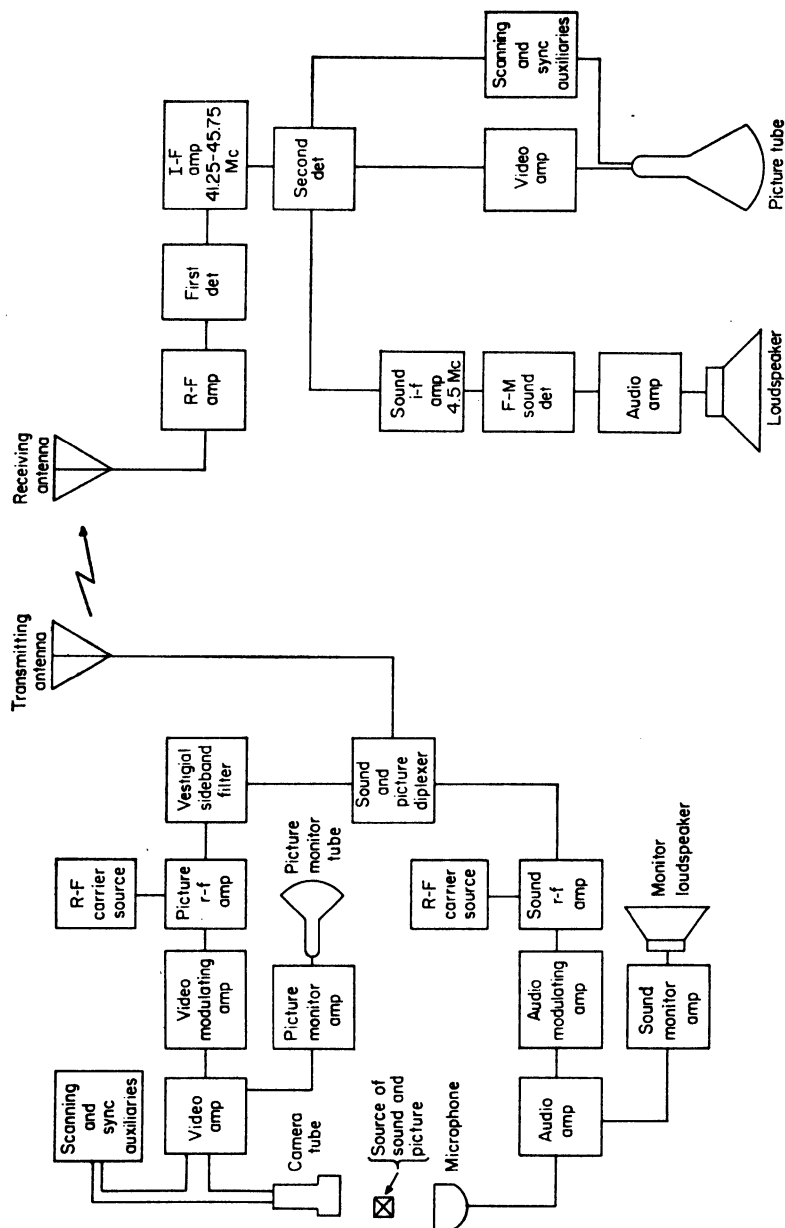



FIG. 1. Elements of television system.

succession. The camera tube, which includes the scanning agent, generates a succession of electrical impulses which correspond with the successive values of brightness discovered by the scanning agent.

At the receiver the scanning process involves setting up an elemental luminous spot of small area which moves synchronously with the scanning agent in the camera tube.

The brightness of this luminous spot is controlled by the electrical impulses transmitted from the camera tube to the receiver. The values of brightness present in the original image are thereby reproduced in their proper positions. The scanning process must be rapid enough so that all the elements of the received image are perceived simultaneously by the eye. This requirement is met if the scanning of the image is completed within the duration of persistence of vision, so that the first element of brightness persists in the eye during the production of all the succeeding elements in the image.

4. Aspect Ratio. The ratio of width (w) to the height (h) of the rectangle actively employed in reproducing the image is known as the *aspect ratio.* In accordance with the standard adopted for motion pictures, in the United States this ratio is given the value

$$\frac{w}{h} = \frac{4}{3} \qquad (1)^*$$

5. Total Number of Lines per Frame. The total number of lines over which the scanning agent passes from the beginning of one complete image to the beginning of the next is known as the total number of lines per frame, n.

The number of lines determines the degree of detail which may be accommodated in the reproduced picture, in the vertical dimension. Hence this number sets an upper limit to the amount of pictorial detail which may be accommodated in that dimension. The number in modern systems is set usually between 400 and 700 lines. According to the FCC standards, n has the value

$$n = 525 \qquad (2)^*$$

The reason for the exact number 525 (see Secs. 6 and 34) is that it is an odd number composed of simple odd factors

$$525 = 3 \times 5 \times 5 \times 7$$

6. Interlaced Scanning. To reduce flicker in the reproduced image, a scanning technique known as *interlacing* is customarily employed, whereby the image is scanned in two groups of lines. The scanning motion in "two-field odd-line" interlaced scanning (the method now universally adopted) is shown in Fig. 2. The scanning agent traverses the area in two series of lines, alternately, passing downward (at left in Fig. 2) from point A to B in the light solid line, following the back-and-forth motions shown by the arrows. The scanning spot then moves upward from point B to C (at right), thence downward again from point C to D on the heavy line, finally upward again from point D to the starting point A, where the motion repeats itself. The scanning agent is active in discovering or reproducing the picture information while traveling over the lines shown solid and is inactive while traveling over the lines shown dashed. The total number of back-and-forth motions made in traversing both series of lines is n. The total number of active lines (shown solid) is n_a. The inactive lines ($n - n_a$)

FIG. 2. Scanning pattern for two-to-one odd-line interlaced scanning.

* Those relationships marked with an asterisk (*) are the official standards prescribed by the FCC for commercial television broadcasting.

are those made by the scanning agent in traveling from the bottom to the top of the picture area (the motions shown at the right). The number of active lines, n_a, for a 525-line image is approximately 490. The exact value of n_a is given by

$$n_a = \frac{n}{1 + b_v} \tag{3}$$

where b_v, the *vertical retrace ratio*, is the ratio between the upward and the downward scanning time, as defined in Sec. 7.

7. Scanning Velocities and Retrace Ratios. The scanning agent is made to traverse the picture area in the interlaced pattern (Fig. 2) by imparting to it horizontal and vertical motions. The spot is displaced horizontally from left to right during a time interval t_h sec, and simultaneously it is displaced vertically downward over a longer time t_v sec. The two motions cause the spot to move slightly downward and to the right until it reaches the right-hand edge of the area. Then the spot is reversed rapidly and is moved to the left in a fraction of the time occupied in moving from left to right. This fraction is the horizontal retrace ratio, b_h. Thus the complete back-and-forth horizontal motion occupies a time of $t_h(1 + b_h)$ sec. The downward motion persists during the succession of back-and-forth motions until the spot reaches the bottom of the area at the end of t_v sec. Thereupon the downward motion is reversed, and the spot is moved upward in a fraction, b_v, of the time consumed in the downward motion, until it reaches the top of the pattern. During the upward motion, several back-and-forth motions are executed, since the horizontal motions are maintained.

The *horizontal retrace ratio* b_h is the ratio of the backward (to the left) scanning time to the forward (to the right) scanning time. Practical values of b_h range from about 0.10 upward to 0.25. The FCC standards set an upper limit to this ratio of 0.15. The *vertical retrace ratio*, b_v, between the upward and the downward times, ranges from about 0.05 to 0.10. The FCC standards set an upper limit to this ratio of 0.08.

8. Interlaced Fields. One set of the two sets of lines in the interlaced pattern is known as an *interlaced field*. Since the total number of lines in the complete frame is an odd number (525), the number of lines per field is a whole number plus one-half (262½). This accounts for the fact that at the end of the first field (Fig. 2) the spot, at point C, has formed but one-half the horizontal motion. The half-line displacement causes the lines in the second field to be displaced vertically from those in the first field by the width of one line. Consequently the lines of one field fall directly between the lines of the preceding field. If the scanning motion is not accurately timed and if the amplitudes of the vertical and horizontal motions are not constant, this interlaced relationship is not maintained, and the lines of one field tend to overlap the lines of the preceding field. This defect is known as *pairing* of the fields. Its effect is to reduce the detail of the reproduced picture in the vertical dimension.

9. Vertical Resolution. The vertical resolution r_v of the scanning pattern is measured by the number of pictorial details or *picture elements* which may be accommodated in the vertical height of the picture area. Each active scanning line is capable of reproducing one such picture element in the vertical direction but, since the picture elements in the image to be transmitted may not fall directly on the scanning lines, the actual number of picture elements which may be accommodated vertically is less than the number of active scanning lines. The vertical resolution r_v is accordingly the number of active scanning lines multiplied by a factor less than 1, the utilization ratio k. The vertical resolution is accordingly

$$r_v = kn_a \qquad \text{elements per picture height} \tag{4}$$

Practical values of the utilization ratio, depending on the perfection of interlacing, range from about 0.6 to 0.9. With $n_a = 490$, r_v accordingly varies from 300 to 440 elements per picture height. The value $r_v = 350$ is commonly attained in properly operated equipment.

10. Horizontal Resolution. The horizontal resolution r_h of the scanning pattern is measured by the number of picture elements which may be accommodated in the

horizontal direction, measured in a width equal to the picture height. The picture height is used as the basis to make the horizontal resolution directly comparable with the vertical resolution. The total number of picture elements accommodated in the picture width is the horizontal resolution multiplied by the aspect ratio.

The value of r_h does not depend on the dimensions of the scanning pattern but rather on the electrical performance of the television system in reproducing rapid changes of voltage whereby the reproducing scanning agent is changed in brilliance as it moves across each line. In terms of the maximum frequency f_{max} in the video range (see Sec. 16), the horizontal resolution is approximately

$$r_h = 84 f_{max} \quad \text{elements per picture height} \tag{5}$$

where f_{max} is expressed in megacycles. This expression assumes transmission at a rate of 30 frames per second and 525 lines. Horizontal resolution of 340 lines (f_{max} = 4.0 Mc) is attainable in properly operated equipment.

11. Resolution Ratio. The ratio of the horizontal resolution to the vertical resolution is the resolution ratio m:

$$m = \frac{r_h}{r_v} = \frac{84 f_{max}}{k n_a} \tag{6}$$

Unity resolution ratio (equal resolution in vertical and horizontal directions) is not essential for good reproduction; *i.e.*, the resolution in one direction may exceed that in the other without wasting the detail in the direction of higher resolution. In present practice the resolution ratio approaches 0.95, depending on the maximum frequency in the video range. For values of f_{max} = 4 Mc, n_a = 485 lines, and k = 0.75, the resolution ratio is 0.925.

FIG. 3. Relationship between horizontal resolution and maximum video frequency.

12. Total Number of Reproducible Picture Elements. A significant figure of merit of the television system is the total number, N, picture elements which may be accommodated in the picture area, *i.e.*, the product of the number of elements vertically r_v times the number horizontally $(w/h)r_h$:

$$N = \left(\frac{w}{h}\right) r_h r_v = \left(\frac{w}{h}\right) (84 f_{max})(k n_a) \tag{7}$$

$$= \left(\frac{w}{h}\right) m k^2 n_a^2$$

For values of $(w/h) = 4/3$, f_{max} = 4.0 Mc, n_a = 485 lines, and k = 0.75 (m = 0.925), the total number is N = 165,000 picture elements. Performance above 200,000 picture elements is exceptional in the present state of the art.

13. Viewing-distance Relationships. The desirable viewing distance of a television image depends on the resolution available. If we assume a visual acuity of 1 min of arc (typical of most normal eyes), two stationary dark picture elements separated by a bright element (Fig. 4) may be barely resolved by the eye at a distance d_c.

$$d_c = \frac{6{,}876 h}{k n_a} \tag{8}$$

and the corresponding ratio of critical viewing distance to picture height is

$$\frac{d_c}{h} = \frac{6{,}876}{k n_a} \tag{9}$$

For a vertical resolution $r_v = kn_a$ of 350 elements per picture height, the foregoing ratio is about 20 times. This is the maximum viewing distance (20 times the picture height), beyond which the eye is unable to resolve the detail actually present in a stationary image. In practice, since the picture elements are not sharply defined and since the act of following motion reduces visual acuity, full detail can rarely be resolved at distances greater than eight times the picture height.

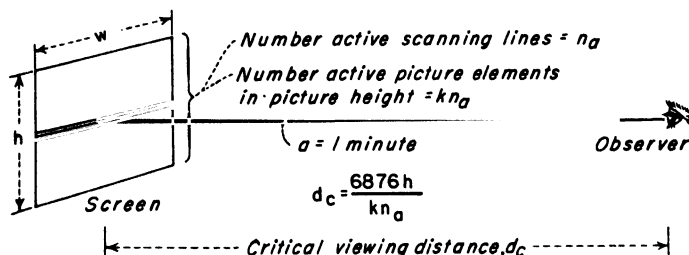



FIG. 4. Critical viewing distance in terms of the dimensions of the scanning pattern.

The minimum viewing distance is determined by the tolerance of the viewer toward the structure of the picture, which becomes increasingly evident as the viewing distance is decreased. Viewing distances shorter than three times the picture height are seldom considered satisfactory. A ratio of 5:1 seems to be typical of viewer habits. Figure 5 shows Eq. (9) for various numbers of scanning lines.

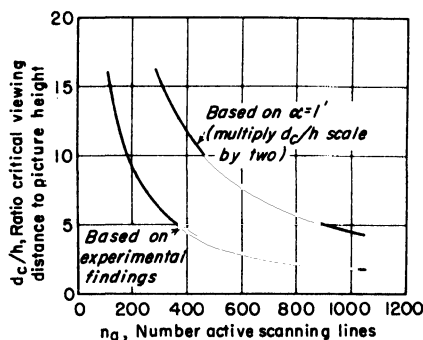



FIG. 5. Viewing distance vs. number of scanning lines, in terms of picture height.

14. Frame-repetition Rate. The rate at which the frames are repeated (frame-repetition rate) depends (1) upon the duration of the persistence of vision of the eye and (2) upon the necessity of reproducing motion in the image in a smooth manner. In motion pictures the standard rate is 24 frames per second, with each frame projected twice, making 48 projection intervals per second. Similar values serve for television. However, since the power-supply frequency for most areas in this country is 60 cps, it is desirable to use a frame-repetition rate f which is a submultiple of the power frequency, *e.g.*, 30 per second (field repetition rate f' of 60 per second). The FCC standards for these items are

$$f = 30 \text{ frames per second} \qquad (10)^*$$

$$f' = 60 \text{ fields per second} \qquad (11)^*$$

15. Rate of Scanning Picture Elements. The maximum rate of scanning picture elements along each line equals the number of elements in the line divided by the time during which the line is scanned. These quantities in turn depend on the horizontal resolution (Sec. 10) and on the number of lines per frame (Sec. 5) and the rate of frame repetition (Sec. 14). The general expression for the maximum rate of scanning picture elements R is

$$R = \frac{w}{h} mfkn^2 \frac{(1 + b_h)}{(1 + b_v)} \quad \text{elements per second} \qquad (12)$$

where the quantities have been defined in the preceding sections. For aspect ratio $w/h = 4/3$, resolution ratio $m = 0.925$, frame-repetition rate $f = 30$ per second, utili-

zation ratio $k = 0.75$, number of lines per frame $n = 525$, horizontal retrace ratio $b_h = 0.20$, and vertical retrace ratio $b_v = 0.08$, the rate of scanning picture elements is approximately $R = 8{,}500{,}000$ elements per second, which is approximately the upper limit of performance of present-day equipment.

16. Maximum Frequency in Video Range. The maximum video frequency generated by the television camera is directly proportional to the rate at which the picture elements are scanned along each line. In deducing a relationship between the scanning rate R (Sec. 15) and the maximum video frequency (v.f.), it is customary to assume that the picture elements are arranged as alternate black and white squares along the scanning line. An ideal scanning agent, scanning such a line, will produce a square wave, as shown in Fig. 6. The upper portion of each square wave represents a black element, the lower portion an adjacent white element. Hence there are two elements per cycle of the wave. The fundamental frequency of the square wave is accordingly one-half as great as the rate of scanning picture elements. The maximum v.f. is then derived from Eq. (12), as

FIG. 6. Wave forms resulting from scanning a checkerboard image. The ideal square wave becomes a sine wave when only the fundamental frequency is transmitted.

$$f_{\max} = \frac{(w/h)mfkn^2}{2} \cdot \frac{(1 + b_h)}{(1 + b_v)} \tag{13}$$

For the conditions cited in Sec. 15, $f_{\max}$ is 4.25 Mc. Table 1 gives other typical values. It should be noted that this frequency is the fundamental of the square wave. The

Table 1. Maximum Video Frequencies for Different Scanning Patterns

No. of scanning lines, n	No. of frames per second, f	Max v.f. for equal vertical and horizontal resolution ($m = 1.00$), cps	Max v.f. for horizontal resolution = 0.9 × vertical resolution ($m = 0.9$), cps
20	16	3,360	3,020
60	16	30,200	27,250
120	24	81,500	163,000
180	24	410,000	369,000
240	24	727,000	653,000
343 (7 × 7 × 7)	30	1,860,000	1,670,000
441 (3 × 3 × 7 × 7)	30	3,060,000	2,800,000
525 (5 × 5 × 7 × 3)	30	4,350,000	3,920,000
1,029 (3 × 7 × 7 × 7)	30	16,650,000	14,800,000

NOTE: Calculation based on $w/h = 4/3$, $b_h = 0.143$, $b_v = 0.083$, $k = 0.75$.

reproducing equipment cannot reproduce a square wave of this frequency. Instead a sine-wave distribution of light is reproduced. This sine wave (Fig. 6) establishes the basic structure of the reproduced image.

17. Scanning Wave Forms. The deflecting forces necessary to produce the linear scanning motions shown in Fig. 2 are *saw-tooth waves*, as shown plotted against time

in Fig. 7. The forward motion in the horizontal direction is produced by a deflecting force linear with time, and the retrace motion by a force which need not be linear but which must have a rate of change high compared with that of the forward force. The ratio of the slopes is equal to the inverse of the horizontal retrace ratio b_h (Sec. 7). The same conditions apply to the deflecting force in the vertical direction, and the ratio of the retrace slope to the forward slope of this wave equals the inverse of the vertical retrace ratio, b_v.

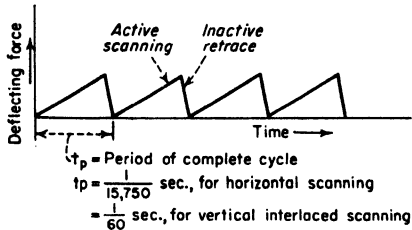



FIG. 7. Saw-tooth waves of deflecting force used to produce the scanning pattern (525-line image).

The scanning wave forms have fundamental frequencies determined by the number of fields per second and by the number of lines per second. In the vertical direction the scanning force must repeat itself at the field repetition rate, $f' = 60$ cps. In the horizontal direction the deflecting force must repeat itself at the line-scanning frequency (525 lines per frame, 30 frames per second), which is the product

$$nf = 525 \times 30 = 15,750 \text{ cps}$$

These values of scanning frequency are the fundamentals of the saw-tooth wave. From 5 to 20 harmonics should be present if the wave form is to approximate the saw-tooth shape sufficiently accurately for scanning purposes. Figure 8 shows the degree of approximation for a saw-tooth wave having a retrace ratio of 0.15 when 5 and 15 harmonics are included. The fifteenth harmonic extends the range up to 900 cps for

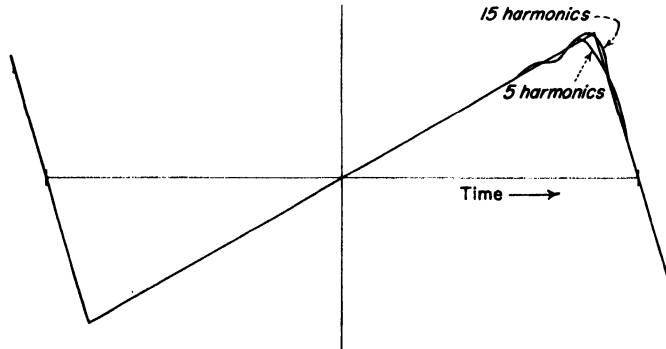



FIG. 8. Ideal saw-tooth wave and approximations resulting from inclusion of 5 and 15 harmonics.

the vertical scanning system, and up to 250,000 cps for the horizontal scanning system. Practical scanning generators are discussed in Sec. 66.

References

Scanning and Image Analysis:

Baldwin, M. W., Jr.: Subjective Sharpness of Simulated Television Images, *Proc. IRE*, **28**, 458, October, 1940.

Bedford, A. V.: Figure of Merit for Television Performance, *RMA Eng.*, **2** (1), 5, November, 1937; also *RCA Rev.*, **3** (1) 36, July, 1938.

Cawein, M.: Television Resolution as a Function of Line Structures, *Proc. IRE*, **33**, 855, December, 1945.

Duke, V. J.: A Method and Equipment for Checking Television Scanning Linearity, *RCA Rev.*, **6**, 190, October, 1941.

Engstrom, E. W.: A Study of Television Image Characteristics, *Proc. IRE*, Part 1, **21**, 1631, December, 1933; Part II, **23**, 295, April, 1935.

Fink, D. G.: Photographic Analysis of Television Images, *Electronics*, **14**, 24, August, 1941.

———: Brightness Distortion in Television, *Proc. IRE*, **29**, 310, June, 1941.

Graham and Reynolds: A Simple Optical Method for the Synthesis and Evaluation of Television Images, *Proc. IRE*, **34**, 18W, January, 1946.
Jesty and Winch: Television Images—Analysis of Their Essential Qualities, *J. Telev. Soc.*, **2**, 316, December, 1937.
Kallmann, H.: The Gradation of Television Images, *Proc. IRE*, **28**, 170, April, 1940.
Kell, R. D., A. V. Bedford, and G. L. Fredendall: Determination of Optimum Numbers of Lines in a Television System, *RCA Rev.*, **5**, 8, July, 1940.
———, ———, and Trainer: Scanning Sequence and Repetition Rate of Television Images, *Proc. IRE*, **24**, 559, April, 1936.
Mertz and Gray: Theory of Scanning and Its Relation to the Transmitted Signal in Telephotography and Television, *Bell System Tech. J.*, **13**, 464, July, 1934.
Wheeler and Loughren: The Fine Structure of Television Images, *Proc. IRE*, **26**, 540, May, 1938.

VIDEO-SIGNAL WAVE FORM

18. Video Signal. The video signal (or "composite video signal") is the succession of electrical impulses transmitted through the television system to convey the information from the scanning agent in the camera to the scanning agent in the receiver. Three direct functions are carried out through the video signal: (1) the transmission of impulses corresponding to the brightnesses of the scanned picture elements, conveyed by the *camera signal;* (2) the blanking of the scanning agent at the receiver during the retrace motions, by the *blanking level* or *pedestal;* and (3) the synchronization of the scanning agents, by the *vertical* and *horizontal synchronization signals.* The first item of the video signal is generated in the camera, the second two in the *synchronization signal generator.* The three items are combined in the *video mixing amplifier.*

19. Envelope of the Modulated Picture-carrier Signal. When the video signal is imposed on a carrier wave, the envelope of the modulated carrier wave constitutes the video-signal wave form. Such a modulated picture carrier and the details of the envelope are shown in Fig. 9.

In the FCC standard video signal (Fig. 9) the carrier amplitude is divided by the black level (blanking level or pedestal) at 75 ± 2.5 per cent of the maximum amplitude. The amplitude region above the black level is called the *infrablack* region and is occupied by the synchronizing signals. Signal levels in this region do not produce light in the received image. The synchronizing signals (Fig. 12) are of two types: (1) horizontal signals for initiating the motion of the scanning agent along each horizontal line and (2) vertical signals for initiating the motion of the scanning agent vertically at the beginning of each field. The peak amplitude of the wave, the height of the synchronizing pulses, and the black-level amplitude are maintained constant throughout each broadcast at the values shown in Fig. 12.

20. Camera Signal. The portion of the carrier envelope extending below the black level is called the *camera signal.* The polarity of transmission in the FCC standard is negative; *i.e.*, an increase in the light on the camera plate results in a decrease in the carrier amplitude, as shown in Fig. 9. The maximum white level is 12.5 ± 2.5 per cent of the maximum carrier level. The black reference level is separated from the blanking level by the *setup* level, which is 75 ± 2.5 per cent of the video range from blanking level to reference white level. The setup level was established as an operating tolerance, the magnitude of which is a compromise between loss of picture-signal amplitude range and clipping of black signal peaks which exceed the blanking level. Intermediate gray tones exist between the reference maximum-white level and the reference maximum-black level. The various levels in the United States composite television signal, expressed as per cent modulation of the picture carrier, are shown in Fig. 9*a*.

A standard wave-form cathode-ray-tube monitor is used in the United States. It employs a scale using blanking level as the zero reference, reference white level as 100 per cent, and synchronizing peak level as about 40 per cent negative. The cover faces of the monitors are inscribed with this scale in conjunction with light horizontal reference lines. This scale is shown in Fig. 9*b*.

The carrier envelope corresponding to the FCC standard is shown in Fig. 9*c*.

The camera signal has two components (Fig. 10): an a-c component, which describes the variations in brightness from the average brightness; and the d-c component,

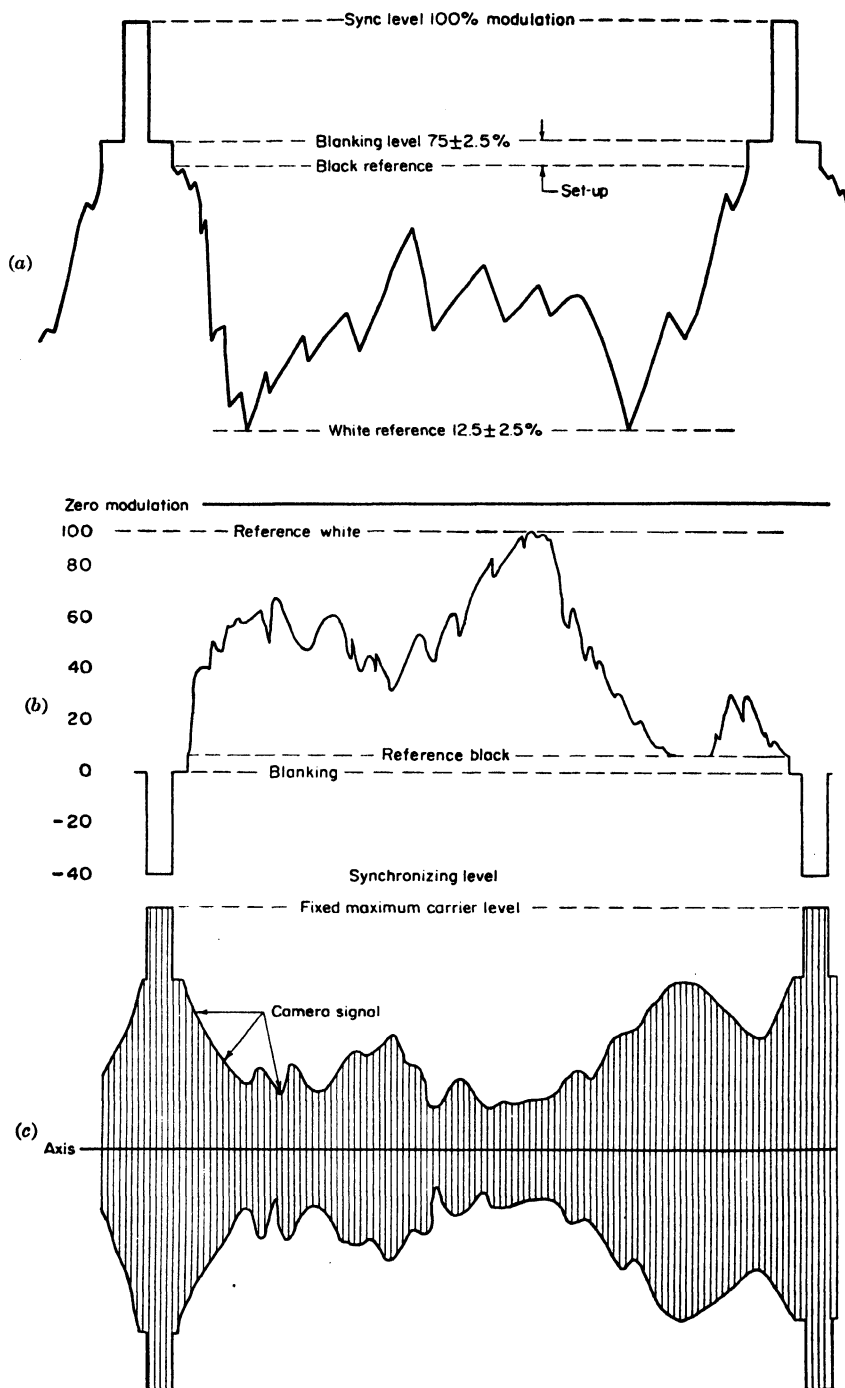



FIG. 9. (*a*) Standard carrier levels for U.S. composite television signals. (*b*) Calibration of cathode-ray-tube wave-form monitor. (*c*) Appearance of carrier envelope.

averaged over the frame-scanning interval ($\frac{1}{30}$ sec), which represents the average or background brightness of the picture. The a-c and d-c components must be capable of being varied independently of each other, so that the same detail may be presented either on a dark background or on a bright background. Variation of the d-c component also permits the screen brightness to be "faded in" or "faded out" at the will of the studio operator. In order that the d-c component be independent of the a-c component, regardless of the changes in wave form, it is necessary that the black level be constant in the carrier envelope, and furthermore that the black level be maintained constant at the control grid of the picture tube (see Sec. 60).

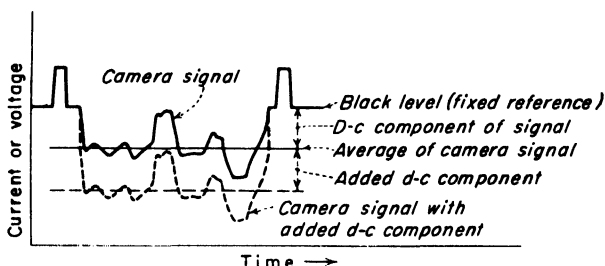



FIG. 10. A-c and d-c components of modulation envelope. Dashed line represents an increase in the background brightness without change in detail.

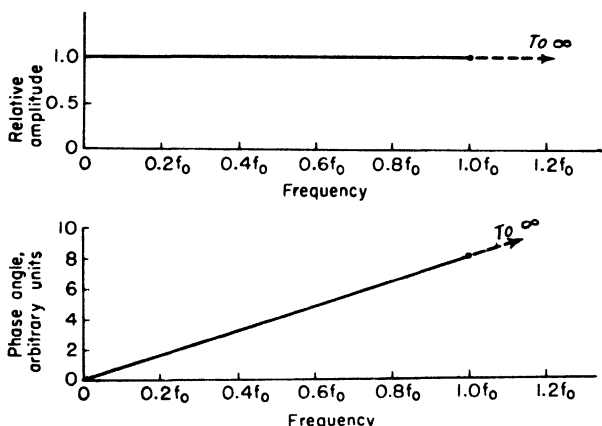



FIG. 11. Ideal phase and amplitude characteristics.

21. Frequency Range in the Video Signal. The maximum frequency in the video range (Sec. 16) results from scanning the finest detail in the image, *i.e.*, from the scanning of adjacent picture elements. The value of f_{max} [Eq. (13)] depends on the rate at which adjacent picture elements are scanned; values up to 4 or 5 Mc are commonly employed in present equipment.

The lowest frequency in the video range, f_{min}, depends on the rate at which the background brightness of the scene changes. Brightness changes which take longer than the duration of a single frame to complete themselves are usually introduced by changes in the d-c component of the signal. Changes that take less than the duration of a single frame are accommodated by video frequencies extending downward to 30 cps (corresponding to the frame-repetition rate of 30 per second). Consequently the significant frequency range in the video signal, based on the FCC standards, is from 30 cps to 4 Mc.

Degrees of picture detail intermediate between the whole frame area and the area of a single picture element are reproduced by frequencies intermediate between 30 cps and 4 Mc. Since such intermediate degrees of detail may be present in any scene, the video-signal transmission system must be equally responsive to all frequencies within these limits.

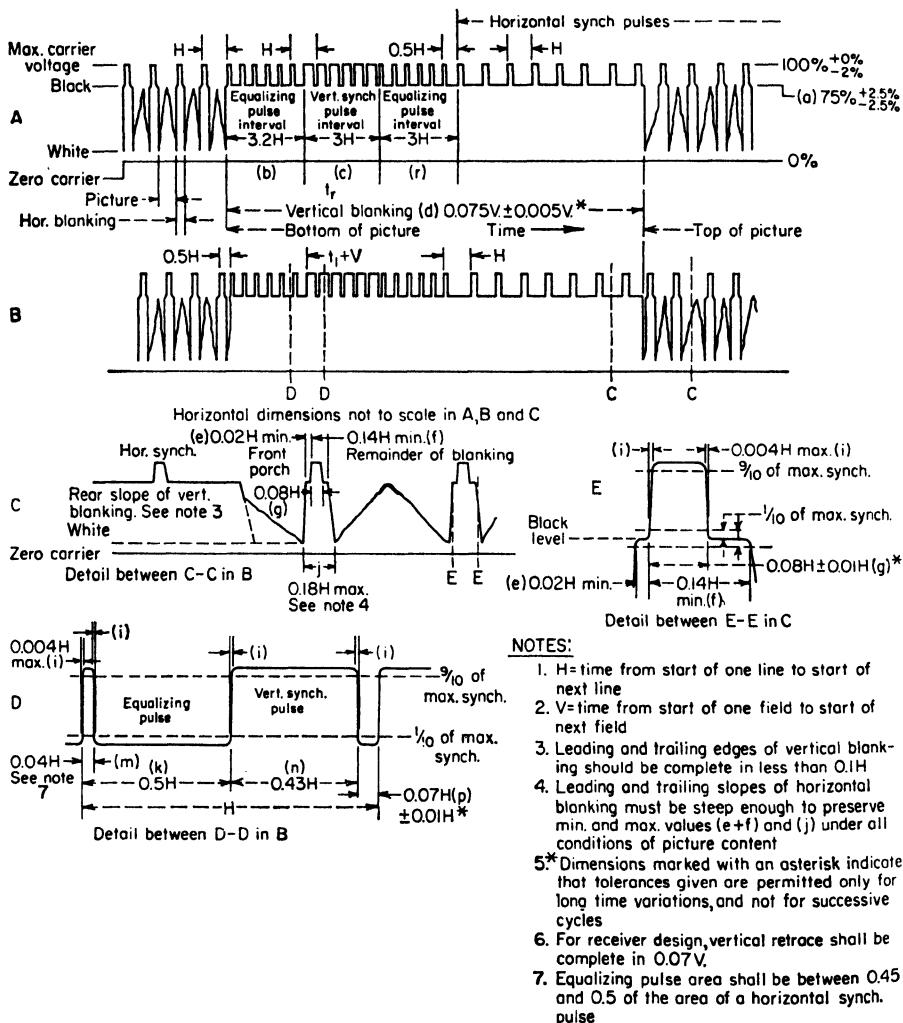



Fig. 12. Dimensions of the FCC standard video-signal wave form.

22. Requirements for Transmission of the Video Wave Form. Fourier analysis of wave forms reveals that any wave form encountered in practice is composed of a number of sine-wave components having specified relative amplitudes and specified relative phases. If the wave form is to be reproduced accurately, the transmission system must be capable of transmitting all such sine-wave components, throughout the v-f range, without altering the relative amplitudes and phases of the components. This

requirement is met if the amplitude-frequency response curve of the transmission system is a horizontal straight line over the v-f range, and if the phase-frequency curve of the transmission system is an oblique straight line passing through the zero-frequency point and extending over the v-f range. The ideal characteristics are shown in Fig. 11.

If the amplitude transmission characteristic is not ideal, the wave form is distorted symmetrically about a vertical axis. If the phase transmission characteristic is not

FIG. 13a. Synchronizing details of Gerber and U.S.S.R. systems. See Fig. 12 for notes.

ideal, the wave form is distorted asymmetrically about a vertical axis. Inadequate h-f response produces improper reproduction of steep vertical changes in the wave form. Inadequate l-f response produces improper reproduction of the flat top portions of the wave which extend over intervals comparable with the period of the low frequencies.

23. Dimensions of Synchronizing Pulses. The dimensions of the sync pulses in the FCC standard wave form are shown in Fig. 12. There are three distinct types of sync pulse. The *horizontal sync pulse* exists on the blanking pulse between the scanning of each line and occupies a duration of 8 per cent of the duration of the line-scanning interval. The *vertical sync pulse* exists on the blanking impulse between the scanning

of successive fields and extends for three times the duration of the line-scanning interval. The *equalizing sync pulses* exist immediately before and after the vertical sync pulse in two groups of six pulses each. The frequency of the equalizing pulses is twice that of the horizontal sync pulses.

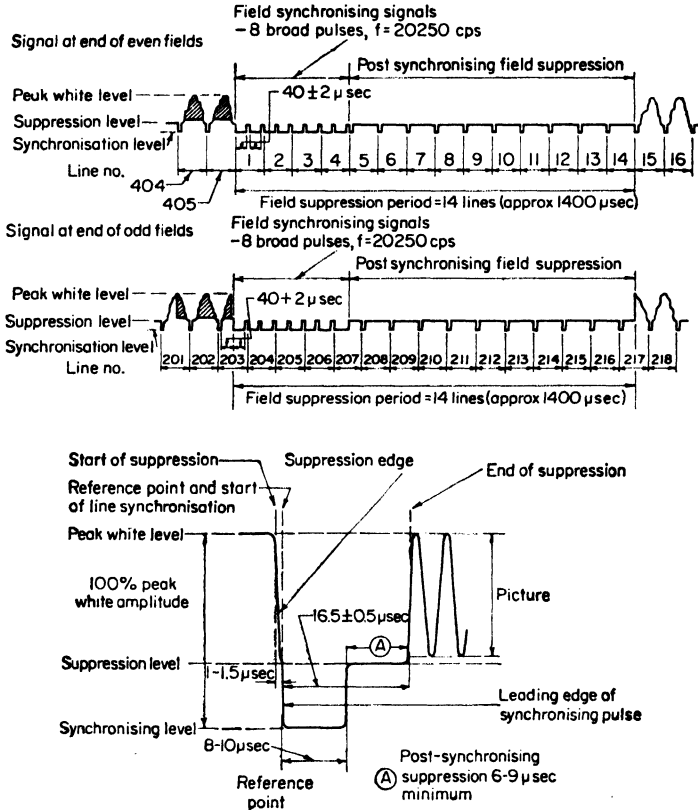



FIG. 13*b*. Synchronizing details of the British system. The hatched part of the signal shown can be occupied either by a suppression pulse up to two lines in length or by picture signal as shown. A small presynchronizing suppression pulse is shown at the opening of the frame suppression. This may be between 0 and 10 μsec in length. The signal is shown in its video form, but synchronizing level corresponds to 0 to 3 per cent. Suppression level 30 per cent ± 3 per cent, and peak white 100 per cent of carrier amplitude. Time of rise of synchronizing (10 per cent to 90 per cent), 0.25 μsec. Time of rise of suppression edge (10 per cent to 90 per cent) should not exceed 1 μsec. Minimum postsynchronizing suppression period 6 μsec. Field-suppression period 14 lines and variation must be such as not to cause visible jitter on picture. Field frequency is tied to frequency of power lines.

The horizontal scanning generators at transmitter and receiver are usually synchronized by the leading edge of the horizontal pulses. Since the synchronizing action must be precise, the leading edge must be sharp. The rise of this edge must complete itself in 0.4 per cent of the line-scanning interval.

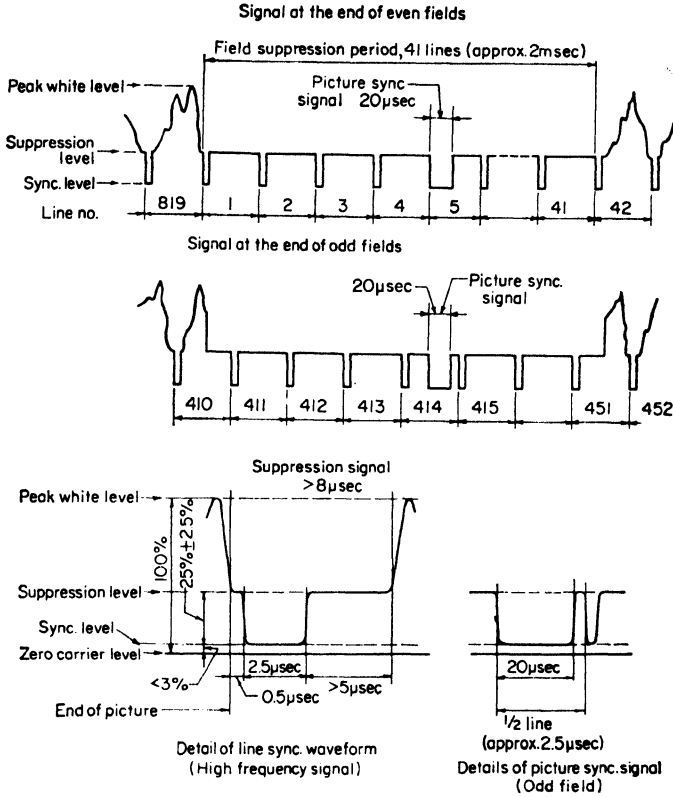



FIG. 13c. Synchronizing details of the French system.

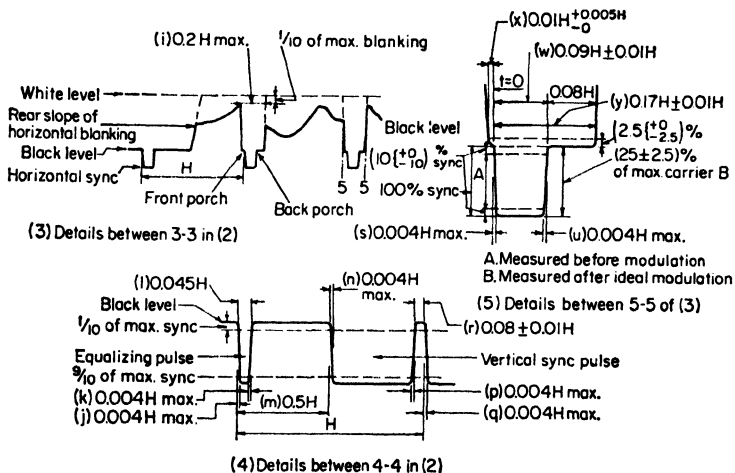



FIG. 13d. Synchronizing details of the Belgian system. See Fig. 13a for (2).

Table 2. World Television Standards Now in Use

	Brazil, Canada, Colombia, Cuba, Dominican Rep., El Salvador, Guatemala, Japan, Mexico, Philippines, Thailand, U.S.A. Uruguay	Australia, Austria, Denmark, Italy, Luxemburg, Netherlands, Norway, Spain, Sweden, Switzerland, Turkey, W. Germany	Czechoslovakia, E. Germany, Hungary, Poland, U.S.S.R.	England, Ireland, New Zealand	France, Monaco, N. Africa, Saar	Argentina, Venezuela	Belgium	Belgium
System identification	U.S.A.	Gerber	U.S.S R.	British	French		Belgian	Belgian
Channel width Mc	6	7	8	5	14	6	7	7
Form of picture modulation	A-m	A-m	A-m	A-m	A-m	A-m	A-m	A-m
Form of sound modulation	F-m ±25	F-m ±50	F-m ±50	A-m	A-m	F-m ±25	A-m	A-m
Picture-modulation polarity	Neg.	Neg	Neg.	Pos.	Pos.	Neg.	Pos.	Pos.
Vestigial side band transmitted	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Full side band transmitted	Upper	Upper	Upper	Lower	Lower	Upper	Upper	Upper
Low channel edge to picture carrier	1 25	1 25	1 25	3 75	11 25	1 25	1 25	1 25
Low channel edge to sound carrier	5 75	6 75	7 75	25	1	5 75	6 75	6 75
Field frequency, monochrome	60	50	50	50	50	50	50	50
Frame frequency	30	25	25	25	25	25	25	25
No. scanning lines	525	625	625	405	819	625	625	819
Line frequency, monochrome	15 750	15 625	15 625	10 125	20,475	15,625	15,625	20,475
Line frequency, color	15 734							
Field frequency, color	59 94							
Color subcarrier frequency, Mc	3 579545							
Aspect ratio	4 3	4 3	4 3	4 3	4 3	4.3	4:3	4:3
Interlacing	2 1	2 1	2 1	2·1	2 1	2·1	2 1	2:1
Audio carrier above picture	4 5	5 5	6 5	−3 5	11 15	4 5	5 5	5 5
Video band width	4 1	5	6	3	10 4	4	5	5

The vertical scanning generators at transmitter and receiver are usually synchronized by the integrated effect of the equalizing and vertical sync pulses which are used to charge a capacitor. The leading edge of the capacitor charge curve (Fig. 45) acts as the synchronizing agent. This leading edge must have precisely the same shape for each vertical pulse. The equalizing pulses are inserted to ensure that this condition is met equally for fields ending on a half line and for fields ending on a whole line.

24. World Television Standards. There are eight recognized sets of television standards in use in the world. Some features are common to all but there are also many differences. In countries which use 50-cps electric power the field frequency is 50 and the frame frequency is 25. In other countries which use 60-cps power the field frequency is 60 and the frame frequency is 30. Television channel widths vary from 5 to 14 Mc. In Europe six different sets of standards are in use.

Table 2 shows important technical specifications of each set of standards and the countries in which they are used.

25. Composite Wave Forms of Other Television Systems. The composite wave forms and dimensions of systems in use which differ from those of the United States

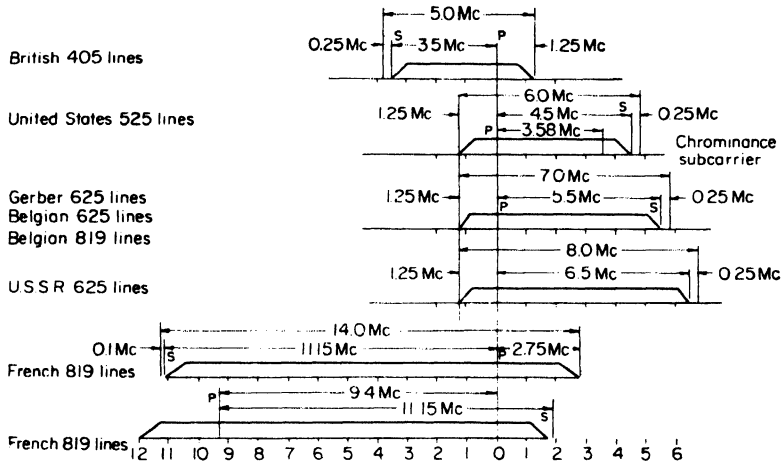



FIG. 14. Channel arrangement for various systems.

system are shown in a series of figures. The Gerber and U.S.S.R. systems are illustrated in Fig. 13*a*. Figure 13*b*, *c*, and *d* shows, respectively, the British, French, and Belgian systems.

The following notes apply to Fig. 13.

1. H = time from start of one line to start of next line.
2. V = time from start of one field to start of next field.
3. Leading and trailing edges of vertical blanking should be complete in less than $0.1H$.
4. Leading and trailing slopes of horizontal blanking must be steep enough to preserve minimum and maximum values of $(x + y)$ and (i) under all conditions of picture content.
5. Dimensions marked with asterisk indicate tolerances and are for long-time variations and not for successive cycles.
6. Equalizing pulses shall be between 0.45 and 0.5 of the area of a horizontal sync pulse.

The manners in which the various systems differ in channel width and location of carriers are illustrated in Fig. 14.

26. Allocation of Television Channels. Frequency allocations for the radio services are the subject of international agreements. The bands assigned for television in the three regions of the world are shown in Table 3. The frequencies in

Table 3. International Frequency Allocations for Television Broadcasting by Blocks

	Band I, Mc	Band III, Mc	Bands IV and V, Mc
Region 1:			
Europe, U.S.S.R. inside and outside Europe, Mongolia, Asia Minor, Africa, and additional out-of-band frequencies for:	41–68	174–216	470–585
France	....	162–174	
Italy	81–88		
Belgium, Germany, Switzerland	....	216–223	
Region 2:			
Western Hemisphere including Hawaii	54–72 76–88	174–216	470–890
Region 3:			
Oceania, Asia (less parts in region 1)	54–68 76–88	174–216	470–585 610–940
Australia	49–56 63–70 85–92 132–146	174–202 209–216	500–855

Television Channels in the United States

Channel No.	Frequency Band, Mc	Channel No.	Frequency Band, Mc
	V.h.f.	40	626–632
2	54–60	41	632–638
3	60–66	42	638–644
4	66–72	43	644–650
5	76–82	44	650–656
6	82–88	45	656–662
7	174–180	46	662–668
8	180–186	47	668–674
9	186–192	48	674–680
10	192–198	49	680–686
11	198–204	50	686–692
12	204–210	51	692–698
13	210–216	52	698–704
		53	704–710
		54	710–716
	U.h.f.	55	716–722
		56	722–728
		57	728–734
14	470–476	58	734–740
15	476–482	59	740–746
16	482–488	60	746–752
17	488–494	61	752–758
18	494–500	62	758–764
19	500–506	63	764–770
20	506–512	64	770–776
21	512–518	65	776–782
22	518–524	66	782–788
23	524–530	67	788–794
24	530–536	68	794–800
25	536–542	69	800–806
26	542–548	70	806–812
27	548–554	71	812–818
28	554–560	72	818–824
29	560–566	73	824–830
30	566–572	74	830–836
31	572–578	75	836–842
32	578–584	76	842–848
33	584–590	77	848–854
34	590–596	78	854–860
35	596–602	79	860–866
36	602–608	80	866–872
37	608–614	81	872–878
38	614–620	82	878–884
39	620–626	83	884–890

bands I, III, IV, and V assigned to the Western Hemisphere have been divided into 6-Mc channels in the United States and each channel is identified by a number, as shown in the table.

27. Standards Conversion. It is desirable to exchange programs among countries using different standards. System differences such as the line frequencies impose difficult problems. One difficulty arises when the camera-tube beam scans a different line structure, producing stroboscopic patterns, or line beating. Another difficulty arises when variations in field frequencies occur between the systems. Still another difficulty arises because the picture-tube display is an intensity-modulated light spot and not a continuous image. If the camera output were responsive to changes in total light flux of the picture tube, it would reflect changes in the picture-tube spot brightness and an unconverted component would be present in the output.

The most convenient method of conversion, and the one employed in Europe, is to use a picture-tube phosphor with sufficiently long persistence time, of the order of one field, to present to the camera a nearly continuous image. Zinc-beryllium silicate phosphor has been used with satisfactory results. The decay characteristic is such that the brightness field is reduced to 28 per cent of its original value during the interval of one field scan. This yields an improvement of 23 db over conventional picture-tube mixed sulfide phosphor in the ratio of converter signal to photosignal. Line beating is minimized by making the picture-tube spot size equal to the vertical distance between lines so that the line structure is rendered practically invisible. This could be accomplished by controlling the focusing circuits but it is more convenient to employ vertical spot wobble, which permits close control of the vertical spot coverage.

The signal to be converted is first applied to a stabilizing amplifier which (1) stabilizes black level and removes interfering signals such as hum, (2) makes available sync pulses at input signal rate for time-locking the display unit, and (3) applies high-frequency preemphasis to compensate for aperture losses in the picture tube. Flyback suppression is applied to permit the picture-display setup to be much beyond the area where flyback lines would normally be visible. An image-orthicon camera, employed for reading, is compensated for aperture losses prior to insertion of blanking.

Programs originated in Paris with the 819-line system were, in the past, converted to a 441-line system for domestic use. The 819-line picture is relayed by microwave radio to Cassel, France, where it is converted to the British 405-line system, following which the signal is again relayed by microwave radio to London for distribution over the British network.

Conversion by this method from the British 405-line picture to the French 819-line picture naturally will impose on the French system the limitations in picture detail of the British system, which inherently is of lower definition capabilities.[1]

GENERATION OF THE VIDEO-SIGNAL WAVE FORM

28. Video-signal Generator. This consists of three essential parts: (1) the camera and its auxiliaries, which generate the camera signal component, (2) the synchronizing signal generator, which times and shapes the vertical and horizontal synchronizing signals and the blanking signals, and (3) the control amplifier, which mixes the camera signals with the synchronizing signals and the blanking signals, forming the composite video signal.

29. Television Cameras. The television camera consists of a housing fitted with an adjustable lens which focuses the scene on the photosensitive plate of the camera tube enclosed within the housing. Also enclosed in the housing is a preamplifier which raises the level of the camera signal (usually to about 0.1 volt peak to peak) so that it can be transmitted over coaxial cable without interference. One or more scanning generators or scanning amplifiers may also be included in the camera housing. The camera is ordinarily mounted on a flexible standard so that it may be moved

[1] Martin, A. V. J., International TV Is Here, *Radio and Television News*, December, 1952. Lord, A. V., Standards Converter for International TV, *Electronics*, August, 1953.

readily, and a universal mounting is provided so that the camera may be directed at any angle. Some form of auxiliary monitor system is also provided to enable the operator to keep the image in focus.

At present there are three important types of electronic camera tubes in use in the United States: (1) the *iconoscope*, which is now used only for film projection, (2) the *image orthicon*, and (3) the *vidicon*. All these tubes make use of the storage principle in which the effect of the incident light is stored as charge across a capacitance element.

30. Iconoscope. A typical iconoscope is shown in Figs. 15 and 16 with its optical and electrical auxiliaries. The image is focused on the mosaic plate, which is a mica

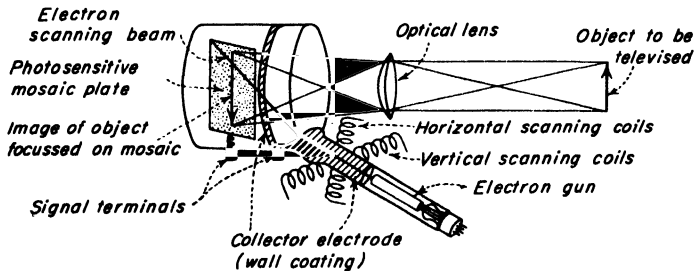



FIG. 15. Optical arrangements used with the iconoscope camera tube.

sheet coated with several million globules of photosensitized silver, insulated from each other and from a metallic coating on the reverse side of the plate. The optical image releases electrons from the mosaic, thereby charging the plate positively with a charge distribution corresponding point for point with the distribution of light in the image. The insulation prevents redistribution of this charge and permits the charge image to increase in magnitude as the light continues to fall on the mosaic.

The mosaic is scanned by a beam of electrons generated in the electron gun in the side arm of the tube. The beam, impinging on the mosaic, releases secondary electrons. The number of secondary electrons released from a given point of the mosaic depends on the potential of that point, which in turn depends on the previous photoelectric emission from that point. Consequently, as the scanning agent passes over the mosaic, it generates a secondary emission current which corresponds to the successive values of brightness in the picture elements. The secondary emission is small for brightly illuminated portions of the mosaic; consequently the output current is "negative" with respect to the illumination responsible for it.

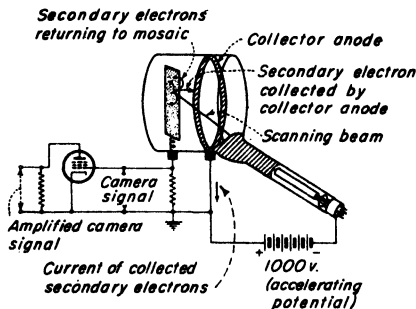



FIG. 16. Iconoscope electrical arrangements.

The secondary emission is collected by a collector anode and conducted through an external coupling resistor back to the metallic signal plate on the back of the mosaic support. The series circuit through which the electron current passes is accordingly composed of the ohmic resistance of the secondary emission path, the coupling resistance, and the capacitance between the signal plate and the group of globules under the scanning agent. No d.c. can flow through the capacitance; hence the output consists simply of the a-c component of the camera signal. The d-c component must be evaluated either by visual observation or by a phototube which integrates the light on the scene. A direct voltage derived from a manual control (or from the phototube in the second case) is inserted in series with the output of the iconoscope.

Since the mosaic is insulated, the current flowing toward or away from it must be zero, when averaged over any extended period of time. The average d-c value of the collected secondary emission must accordingly be replaced by electrons from the scanning beam.

Only a part of the secondary emission is collected from the mosaic. The remainder, falling back on the mosaic, sets up a distribution of charge which, when scanned, produces a spurious signal whose effect is to produce an unevenness in the background shading of the reproduced picture. This spurious signal ("dark-spot signal") must be compensated by a *shading-correction* signal generator (Sec. 38).

The color response of the typical iconoscope mosaic (when the mosaic has been silver sensitized) is very similar to that of the usual panchromatic negative-film emulsion used in motion pictures.

The sensitivity of modern iconoscopes under optimum conditions varies from about 1 mv per millilumen per sq cm illumination on the mosaic surface (low values of illumination) to about 0.25 mv (at higher illumination). The log-log curve between input illumination and output voltage of the typical iconoscope is characterized by an average slope (Sec. 71) of about 0.7. The output voltages may be increased by increasing the current used in the scanning beam, as well as by illuminating the interior of the tube envelope by a bias light.

31. Image Orthicon. The image-orthicon type of camera tube is almost universally used for live studio pickup and for televising outside events because of its great

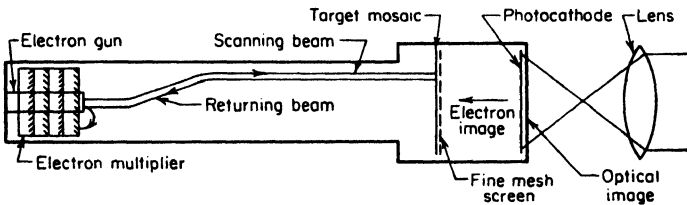



Fig. 17. Image-orthicon construction.

sensitivity, excellent gray-scale reproduction, and over-all merit as a transducer. A diagram of it is shown in Fig. 17. The optical image is focused on the *photocathode*, just inside the glass envelope of the tube. This is a continuous surface of photosensitive material which emits electrons from its rear face in a distribution proportional to the lights and shadows of the image.

The photocathode is several hundred volts negative with respect to the *target electrode* (mosaic), which is located parallel to the photocathode and about 1½ in. distant. The electron image formed at the photocathode is accordingly drawn to the target. There the impinging electrons give rise to secondary emission, the secondaries being collected by a fine-mesh screen located adjacent to the surface of the target. The target thereby becomes charged with a distribution proportional to the brightnesses of the optical image, and the value of this charge is several times greater than the impinging electron image, by virtue of the electron multiplication which occurs in the act of secondary emission at the target surface.

The target is made of very thin glass having low resistivity. The lateral resistance is sufficient to preserve the charge configuration on the target for the duration of the frame interval. The target is scanned on its rear surface by a low-velocity beam of electrons in a manner identical to the scanning in the standard orthicon. Electrons are collected from the scanning beam by virtue of the positive-charge image on the reverse side of the target, and the variations thus imposed on the returning beam of scanning electrons constitute the video signal.

The electrons absorbed on the rear face of the target move through the target, by virtue of the low-resistance path offered by the glass. The charge image trapped on the forward face is thus neutralized and leaves the target prepared for the next scanning cycle.

The returning beam impinges on a target structure surrounding the electron gun, which leads to an electron multiplier. The amplified video current is collected at the output electrode of this structure and passed through the coupling resistor.

The image orthicon has a sensitivity several hundred times greater than that of the iconoscope, the limit of sensitivity being shot noise in the electron emission at the photocathode and at the target electrode. At threshold values of light, the image-orthicon exposure index exceeds manyfold that of the fastest photographic film; a recognizable image may be televised with it in the light of a match. The video signal is linearly related to the incident light up to a limiting value, at which a sharp saturation occurs. This permits the camera to be exposed to very widely varying levels of light without blocking or desensitizing the mosaic. The camera is very successfully used under marginal conditions of light like those encountered in televising outside events.

32. The Vidicon. The RCA vidicon is a small television-camera tube widely used for industrial puposes; for cameras for motion-picture film, transparencies, and

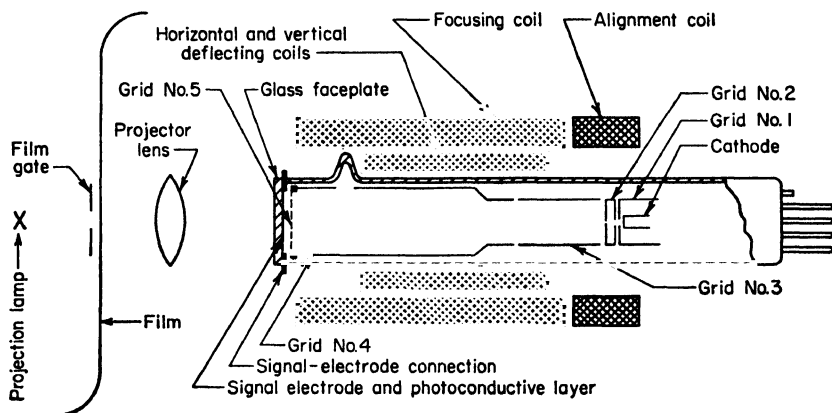



FIG. 18. General construction of the vidicon.

opaques; and in extremely small, compact, and portable cameras. Various models vary in size from 6¼ by 1.125 in. for 600-line-resolution high-quality film cameras to experimental models only about 3 by ½ in. having somewhat lower resolution. The construction of a vidicon tube is illustrated in Fig. 18. For film cameras the tube has a sensitivity permitting televising of motion-picture film with an average highlight illumination of 100 to 300 ft-candles on highlights on the face of the tube. For opaques and transparencies a constant illumination of 10 ft-candles on the face of the tube is adequate. The spectral response is similar to that of the human eye and the tube is excellent for use with film. Its small size and simplicity are added advantages in the design of film cameras and auxiliaries.

The tube (Fig. 19) has a transparent conducting film on the inner surface of the faceplate; a light-sensitive element consisting of a thin layer of photoconductive material deposited on the signal electrode; a fine-mesh screen (grid No. 4) located adjacent to the photoconductive layer; a focusing electrode (grid No. 3) connected to grid No. 4; and an electron gun for producing a beam of electrons.

Each element of the photoconductive layer is an insulator in the dark but becomes slightly conductive when it is illuminated and acts like a leaky capacitor having one plate at the fixed positive potential of the signal electrode and the other floating. When light from the scene being televised is picked up by an optical-lens system and focused on the photoconductive-layer surface next to the faceplate, each illuminated layer element conducts slightly depending on the amount of illumination on the element and thus causes the potential of its opposite surface (on the gun side) to rise

in less than the time of one frame toward that of the signal-electrode potential. Hence, there appears on the gun side of the entire layer surface a positive potential pattern, composed of the various element potentials, corresponding to the pattern of light from the scene imaged on the opposite surface of the layer.

The gun side of the photoconductive layer is scanned by a low-velocity electron beam produced by the electron gun. This gun contains a thermionic cathode, a control grid (grid No. 1), and an accelerating grid (grid No. 2). The beam is focused at the surface of the photoconductive layer by the combined action of the uniform magnetic field of an external coil or permanent magnet and the electrostatic field of grid No. 3. Grid No. 4 serves to provide a uniform decelerating field between itself and the photoconductive layer so that the electron beam will approach the layer in a direction perpendicular to it—a condition necessary for driving the surface to cathode potential. The beam electrons approach the layer at low velocity because of the low operating potential of the signal electrode.

When the gun side of the photoconductive layer with its positive-potential pattern is scanned by the electron beam, electrons are deposited from the beam in sufficient quantities until the surface potential is reduced to that of the cathode, and thereafter are turned back to form a return beam which is not utilized in this tube. Deposition of electrons on the scanned surface of any particular element of the layer causes a change in the difference of potential between the two surfaces of the element. When the two surfaces of the element, which in effect is a charged capacitor, are connected through the external signal-electrode circuit and the scanning beam, a capacitive current is produced and constitutes the video signal. The magnitude of the current is proportional to the surface potential of the element being scanned and to the rate of scan. The video-signal current is then used to develop a signal output voltage across a load resistor. The signal polarity is such that, for highlights in the film, transparency, or opaque, the grid of the first video-amplifier tube swings in a negative direction.

Alignment of the beam is accomplished by a transverse magnetic field produced by external coils located at the base end of the focusing coil.

Deflection of the beam is accomplished by transverse magnetic fields produced by external deflecting coils.

Figure 19 shows a photograph of the type of vidicon used for high-quality film cameras.

FIG. 19. Actual appearance of the vidicon.

33. Preamplifier. A typical preamplifier for use with an iconoscope is shown in Fig. 20. To preserve a high signal-to-noise ratio in the first stage, an effective value of about 300,000 ohms is used as the coupling resistance, with a shunt capacitance of about 8 $\mu\mu$f. The poor h-f response incident to this combination is compensated in the third stage, which employs a bifilar winding (L_2) to remove the effect of the impedance in the power supply. The second and fourth stages are conventional video-amplifier stages with flat response to 5 Mc. The output stage is a cathode-coupled stage having less than unity gain and presenting an output impedance which matches the characteristic impedance (65 ohms) of the coaxial cable. The camera signal is sent over this cable to the control amplifier for mixing with the synchronization impulses. The output of the shading-correction generator (Sec. 38) is inserted directly in series with the signal plate of the mosaic.

34. Synchronization Signal Generator. *Timing Unit.* The sync pulses (Fig. 12) must be properly timed and properly shaped. A typical timing unit is shown in Fig. 22. The unit produces two outputs at 60 and 15,750 cps (for the FCC standard

video signal). The 60-cps output is derived from the basic 15,750-cps oscillation by frequency multiplication and division; multiplication to 31,500 cps and division in four steps of 7 (to 4,500 cps), 5 (to 900 cps), 5 (to 180 cps), and 3 (to 60 cps). Frequency multiplication is carried out in a frequency-converter tube, the divisions occurring in multivibrators isolated by buffer stages.

FIG. 20. Circuit diagram of an iconoscope preamplifier.

FIG. 21. Low-noise cascode vidicon preamplifier having 8-Mc band width.

The locally generated 60-cps signal is then compared with the 60-cps voltage of the power system by feeding the two sources to a discriminator diode which develops a d-c voltage proportional to the amount and direction of the phase difference between the two sources. This d.c. is used in an a-f-c circuit to correct the frequency of the basic 15,750-cps oscillator from which the locally generated 60 cps is derived. In

FIG. 22. Typical synchronization-signal-generator timing unit for 525-line image.

this way the 15,750- and 60-cps outputs are maintained in synchronous relationship with each other and with the frequency of the power system.

35. Synchronization Signal Generator. *Horizontal Shaping Unit.* Figure 23 shows one portion of the shaping unit of the sync signal generator, the horizontal shaping unit. This unit accepts the 15,750-cps output of the timing unit and produces from it the several wave forms required for the EIA standard signal (Fig. 12). The upper chain of tubes produces the horizontal sync pulses at 15,750 cps, the successive tubes being employed to obtain the required duration, shape, and steepness of front required for these pulses. The middle chain of tubes produces the serrated vertical sync pulses continuously at 31,500 cps. The bottom chain produces equalizing

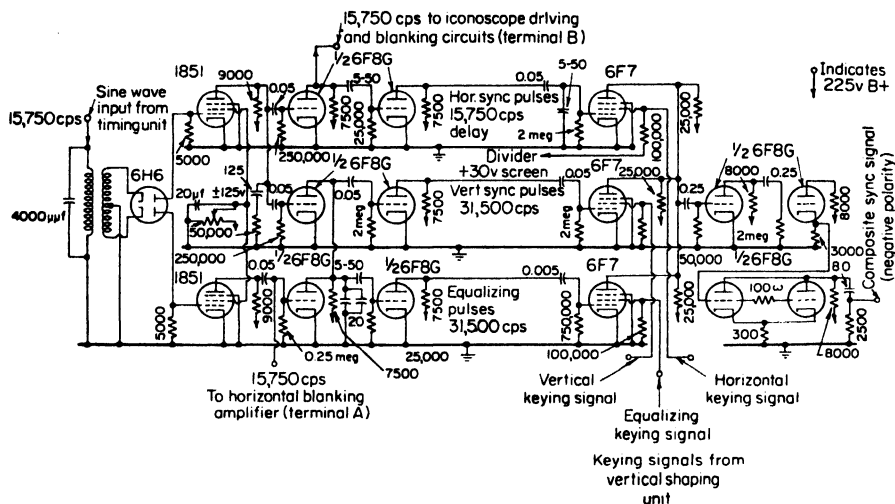



FIG. 23. Typical horizontal shaping unit of a synchronization signal generator; 525-line image.

pulses at 31,500 cps. All three types of pulse are produced continuously in this portion of the generator. They are interspersed in the proper order (Fig. 12) by the action of keying signals in the 6F7 tubes at the right of each chain. The interspersed signal (composite sync signal) is then amplified by the stages at the extreme right and is applied to the control amplifier for mixing with the camera signal.

36. Synchronization Signal Generator. *Vertical Shaping Unit.* The vertical shaping unit (Fig. 24) has the function of producing so-called *keying signals* of 60 cps. These keying signals are applied to the screen grids of the keying tubes in the horizontal shaping unit. The action of the keying signals is to allow to pass, or to block, the sync signals passing through the keying tubes. Thus a negative keying signal is required to block the passage of the horizontal sync signals during the vertical blanking period (Fig. 12); a positive keying signal is required to allow the serrated vertical sync pulse to pass at the proper time during the field blanking interval; and a two-part positive keying signal is needed to allow the equalizing pulses to pass immediately before and after the vertical sync pulses. The shape of the keying signals and the synthesis of the composite sync signal are shown in Fig. 25. The vertical shaping unit accepts the 60-cps output of the timing unit and forms the required keying signals by several chains of shaping tubes which introduce the necessary wave shaping and delaying actions.

The vertical shaping unit also provides blanking signals which are applied to the control amplifier to introduce the black level during the retrace periods. Two sets

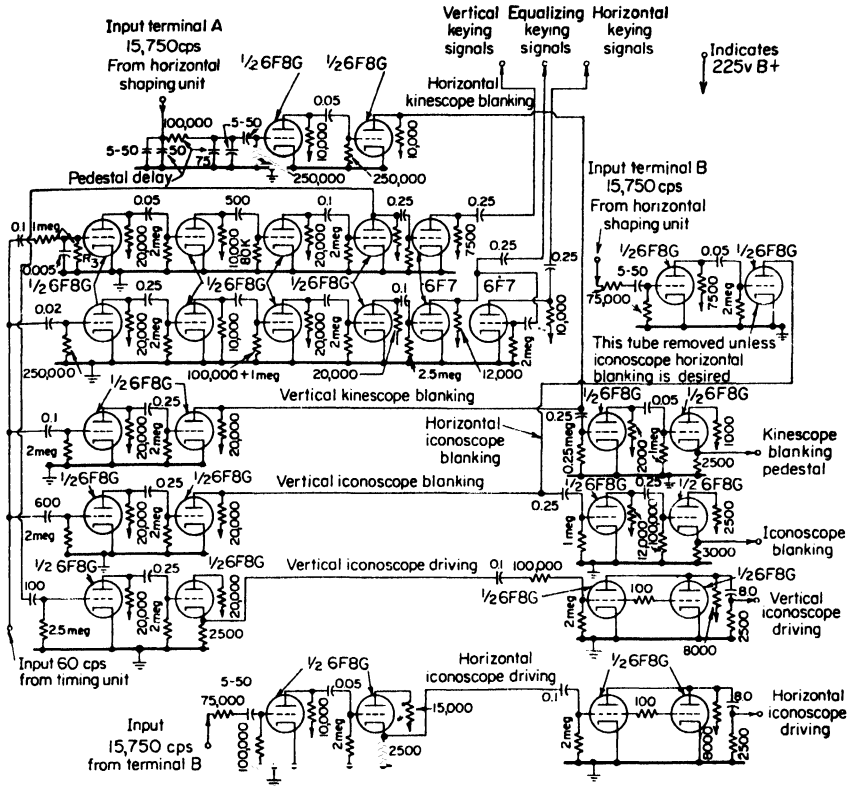



FIG. 24. Typical vertical shaping portion of synchronization signal generator for 525-line image.

of blanking signals are generated: one for the composite video signal and another, of somewhat shorter duration, for the control of the scanning beam in the camera tube. Each group of blanking signals consists of horizontal and vertical square waves recurring at 15,750 and 60 cps, respectively.

The camera-tube scanning generators are controlled by vertical and horizontal driving impulses, which are somewhat narrower and sharper than the corresponding sync pulses in the composite video signal. These driving pulses are formed from 15,750- and 60-cps signals derived from the horizontal and vertical shaping units, respectively.

FIG. 25. Function of the keying signals in interspersing the components of the composite synchronization signal.

37. Mixing Amplifier. The mixing amplifier (Fig. 26) has three input terminals which accept the camera signal from the camera preamplifier, the blanking signals, and the composite sync signals from the sync generator. The camera signal and blanking are first combined by means of two amplifier tubes feeding a common load resistor, across which the "semicomposite" signal appears. The d-c component of the video signal is controlled by varying the bias on the blanking-signal amplifier tube, thus controlling the amplitude relationship between the average of the camera-signal component and the blanking level.

The composite sync signal is similarly added to the camera and blanking components in two amplifier stages feeding a common load resistor, across which the composite video signal appears. Bias controls across these tubes control the relative amplitude

FIG. 26. Typical mixing video amplifier for combining camera signal, blanking signals, and composite sync signals.

of the camera and sync signal amplitudes, thus allowing the establishment of the 75–25 per cent relationship demanded by the standard signal.

38. Shading-correction Generator. The shading-correction generator is a device for producing wave shapes of saw-tooth, sine, and parabolic shape at vertical scanning and horizontal scanning rates (60 and 15,750 cps, respectively) in synchronism with the scanning motion. These wave shapes, controlled as to amplitude, phase, and polarity, are introduced in the preamplifier (Fig. 20) to compensate the spurious shading signal generated in the iconoscope. A form of shading-correction generator is shown in Fig. 27. The horizontal saw-tooth generator used to deflect the beam in the iconoscope tube is used directly to produce saw tooths of controlled amplitude and polarity, as well as 15,750- and 31,500-cps sine waves of controllable amplitude, phase, and polarity. Similarly the output of the vertical saw-tooth generators is used to produce similar signals. Reversal of polarity is provided by an amplifier stage. The switches control the polarity, resistors R_1, R_3, R_5, R_6, R_8, and R_{10} control amplitude, and resistors R_2, R_4, R_7, and R_9 control the phase. Methods of producing saw-tooth waves of controllable phase are also available, using "clipped-off" portions of the basic saw-tooth waves. The shading-correction-generator controls are manipulated man-

FIG. 27. Typical shading-correction signal generator.

ually to correct for the observed defects of shading in the image as viewed on the monitor picture tube.

References

Video Signal and Its Generation, Including Camera Tubes:

Applegarth, A. R.: Synchronizing Generators for Electronic Television, *Proc. IRE*, **34**, 128W, March, 1946.

Barco, A. A.: Iconoscope Preamplifier, Report LB-448 of the RCA License Laboratory. Information made available by special permission. See also *RCA Rev.*, **4** (1), 89, July, 1939.

Iams, Janes, and Hickok: The Brightness of Outdoor Scenes and Its Relation to Television Transmission, *Proc. IRE*, **25**, 1034, August, 1937.

——— and Rose: Television Pick-up Tubes Employing Cathode-ray Beam Scanning, *Proc. IRE*, **25**, 1048, August, 1937.

——— and ———: Television Pickup Tubes Using Low-velocity Electron Beam Scanning, *Proc. IRE*, **27**, 547, September, 1939.

Janes and Hickok: Recent Improvements in the Design and Characteristics of the Iconoscope, *Proc. IRE*, **27**, 535, September, 1939.

Kell, R. D., and G. C. Sziklai: Image Orthicon Camera, *RCA Rev.*, **7**, 67, March, 1946.

McIlwain, Knox: Survey of Television Pickup Devices, *J. Applied Phys.*, **10**, 432, July, 1939.

Maloff, I. G.: Gamma and Range in Television, *RCA Rev.*, **3** (4), 409, April, 1939.

Montfort and Somers: Measurements of the Slope and Duration of Television Synchronizing Pulses, *RCA Rev.*, **6**, 370, January, 1942.

Rose, A.: The Relative Sensitivities of Television Pickup Tubes, Photographic Film, and the Human Eye, *Proc. IRE*, **30**, 293, June, 1942.

———, Weimer, and Law: The Image Orthicon—A Sensitive Television Pickup Tube, *Proc. IRE*, **34**, 424, July, 1946.

Zworykin, V. K.: The Iconoscope, a New Version of the Electric Eye, *Proc. IRE*, **22**, 16, January, 1934.

———, Morton, and Flory: Theory and Performance of the Iconoscope, *Proc. IRE*, **25**, 1071, August, 1937.

CCIR Reports.

VIDEO AMPLIFICATION

39. Requirements for Video Amplification. The system must transmit all sine-wave components within the video range, *e.g.*, 30 cps to 4 Mc, without amplitude discrimination and without phase discrimination. The gain G of a pentode amplifier stage (plate resistance large compared with the load resistance) is

$$G = g_m Z_0 \tag{14}$$

where g_m is the grid-plate transconductance of the tube, and Z_0 is the output impedance of the coupling connection between the stage and the following transducer. Over

the video range g_m is independent of frequency; hence the amplitude and phase responses of the amplifier are determined solely by Z_0.

In video amplifiers, Z_0 consists of R, L, and C components so proportioned as to display a constant magnitude of impedance and a phase angle proportional to frequency over the video range. The lower frequency limit over which these conditions may be met is determined by the series impedance of the coupling capacitor, whereas the h-f limit is determined by the shunt impedance of the capacitance existing in shunt across the coupling connection. The gain within these limits depends on g_m and on the value of the load resistor, since this is the principal component of Z_0 within the video range.

40. H-f Compensation. To minimize the effect of the shunt capacitance, it is usual in video amplifiers to insert a small inductance either in series with the load resistor (shunt peaking), in series with the coupling connection (series peaking), or a combination of the two (shunt-series peaking). The inductance is used to form a resonant circuit with the shunt capacitance at a frequency above the upper limit of the required v-f range, and the rising resonance characteristic is used to counteract the falling off of the Z_0 value at the upper frequency limit. The load resistor must similarly be chosen in terms of the total shunt capacitance, so that the gain in the mid-frequency range (where reactive effects are not prominent) will be the same as at the upper limit (where reactive effects are predominant).

In all cases of h-f compensation the basic factor is the total shunt capacitance C_t associated with the coupling connection

$$C_t = C_{gk} + C_{pk} + C_{gp}(G + 1) + C_{\text{stray}} \tag{15}$$

where C_{pk} = output tube capacitance
C_{gk} = input capacitance of following tube
C_{gp} = grid-plate capacitance of following tube
G = stage gain of following stage
C_{stray} = total shunt capacitance due to wiring, tube sockets, terminals, etc.

In pentode amplifiers, C_{gp} may ordinarily be neglected.

41. Shunt-peaking Compensation. The most widely used h-f compensation scheme (Fig. 28) is known as *shunt peaking*, because the resonating (peaking) inductance L_0 is in shunt across the shunt capacitance C_t. The design values of L_0 and R_0 (the load resistor) are based on the shunt capacitance C_t, on the maximum required frequency in the video range $f_{\max}$, and on two design constants k_L and k_R which relate the impedance of L_0 and R_0, respectively, to the impedance of C_t at the maximum frequency $f_{\max}$.

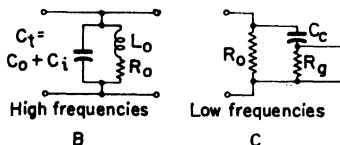



Fig. 28. H-f compensation by the shunt-peaking method, with equivalent circuits for high and low frequencies.

$$k_R = \frac{R_0}{1/(2\pi f_{\max} C_t)} \tag{16}$$

$$k_L = \frac{2\pi f_{\max} L_0}{1/(2\pi f_{\max} C_t)} \tag{17}$$

The values of k_R range from 0.8 to 1.0; most designs are based on $k_R = 1$; *i.e.*, the load resistance is made equal to the impedance of C_t at the maximum v.f. The values of k_L range from 0.3 to 0.7, but most designs are based on $k_L = 0.5$; *i.e.*, impedance of the inductance L_0 is made one-half as great as the impedance of C_t at $f_{\max}$. This is equivalent to making the resonant frequency between L_0 and C_t equal to 1.41 times $f_{\max}$.

On the assumption that $k_R = 1.0$, the gain at frequency f of the shunt-compensated video amplifier is

$$G = \frac{g_m R_0\{1 - j[k_L^2(f/f_{\max})^3 + (1 - k_L)(f/f_{\max})]\}}{(f/f_{\max})^2 + [k_L(f/f_{\max})^2 - 1]^2} \tag{18}$$

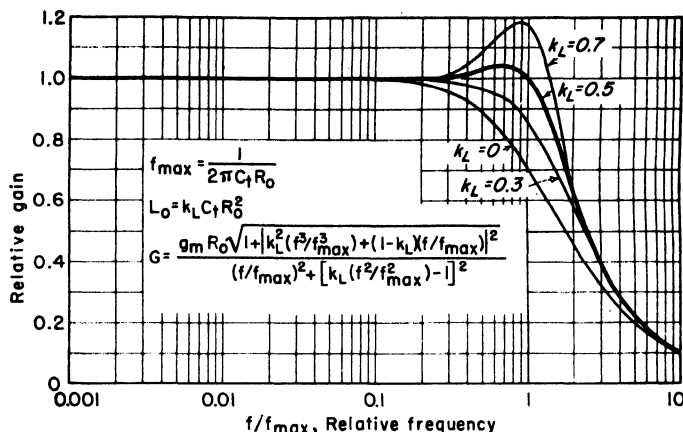



Fig. 29. Gain of a shunt-compensated video amplifier.

The absolute magnitude of this equation is plotted in Fig. 29, and its phase angle in Fig. 30, for several values of k_L.

Simplified design equations for shunt peaking ($k_R = 1$ and $k_L = 0.5$) are:

$$R_0 = \frac{1}{2\pi f_{max} C_t} \tag{19}$$

$$L_0 = 0.5 C_t R_0^2 \tag{20}$$

Typical values of R_0 are 2,000 to 4,000 ohms and of L_0 are 50 to 100 μh.

42. Series-peaking Compensation. The compensation in Fig. 31 has an advantage over the shunt-peaking system in that the inductance L_c isolates the effects of the output and input capacitances C_0 and C_i, whereas in the shunt-peaking systems, C_0 and C_i are directly additive. Since C_0 is less than C_t, for a given h-f limit R_0 may be made correspondingly larger; hence the gain of the stage is increased. On the assumption that $C_i/C_0 = 2$ (usually assumed condition), the design equations for R_0 and L_c are as follows:

Fig. 30. Phase angle introduced by a shunt-compensated video amplifier.

$$R_0 = \frac{1.5}{2\pi f_{max}(C_0 + C_i)} \tag{21}$$

$$L_c = 0.67 C_t R_0^2 \tag{22}$$

With these values the gain is uniform up to f_{max}, and its value is 50 per cent greater than the gain of the shunt-compensated stage with the same values of C_0, C_i, and C_t, provided $C_i/C_0 = 2$.

Fig. 31. Series-peaking system of h-f compensation.

43. Shunt-series-peaking Compensation. The combination of shunt and series peaking (shown in Fig. 32) allows still higher gain by combining the virtues of both connections. Assuming $C_i/C_0 = 2$, the design equations are:

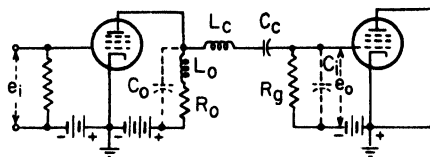



FIG. 32. Combined shunt- and series-peaking systems.

$$R_0 = \frac{1.8}{2\pi f_{\max} C_t} \quad (23)$$

$$L_0 = 0.12 C_t R_0^2 \quad (24)$$

$$L_c = 0.52 C_t R_0^2 \quad (25)$$

The stage displays up to $f_{\max}$ uniform gain, which is 80 per cent greater than that of the simple shunt-peaking stage. The relative merits and design factors of the three methods of h-f compensation are shown in Table 4.

Table 4. High-frequency Compensation Systems

Type	R_0	L_0	L_c	Relative gain at $f_{\max}$	Variation in time delay, sec up to $f_{\max}$ cps
Uncompensated	$1/(2\pi f_{\max} C_t)$	....	....	0.707	$0.035/f_{\max}$
Shunt	$1/(2\pi f_{\max} C_t)$	$0.5 C_t R_0^2$	....	1.0	$0.023/f_{\max}$
Series ($C_i/C_0 = 2$)	$1.5/(2\pi f_{\max} C_t)$	....	$0.67 C_t R_0^2$	1.5	$0.0113/f_{\max}$
Shunt-series ($C_i/C_0 = 2$)	$1.8/(2\pi f_{\max} C_t)$	$0.12 C_t R_0^2$	$0.52 C_t R_0^2$	1.8	$0.015/f_{\max}$

44. L-f Compensation. The amplitude response of conventional RC-coupled amplifier stages at low frequencies is usually satisfactory, but the phase response at low frequencies is troublesome.

The phase angle introduced by C_c and R_g (Fig. 33) is sufficient to prevent proper reproduction of square waves of 30 or 60 cps fundamental frequency, unless very large values of C_c and R_g are employed. Large values of C_c introduce shunt capacitance to ground, and large values of R_g introduce grid-current difficulties. Large values of C_cR_g may induce relaxation oscillations. Accordingly it is usual to compensate the effect of the time constant C_cR_g by the introduction of a filter R_FC_F shown in Fig. 33.

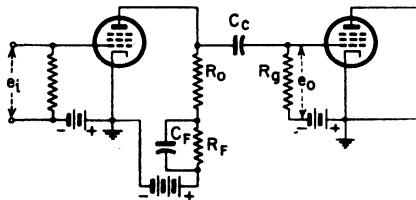



FIG. 33. RC l-f compensation.

$$\frac{C_F R_0 R_F}{R_0 + R_F} = C_c R_g \quad (26)$$

When this condition is met, the gain at low frequencies is

$$G = \frac{(f/f_F) g_m R_0}{f/f_F - j} \quad (27)$$

where G = gain at frequency f
$f_F = 1/(2\pi C_F R_F)$
$j = \sqrt{-1}$

The amplitude and phase of Eq. (27) are shown in Figs. 34 and 35. Values of R_FC_F from 0.15 to 0.5 should be used to keep the point of zero-phase shift below 30 cps, as indicated in Fig. 35.

45. Cathode-coupled Stage. For many purposes a video-amplifier stage displaying low output impedance is necessary (to match the impedance of coaxial cables and to permit the stage to feed many high-impedance sources at once). The cathode-coupled stage (Fig. 36) is commonly used for this purpose. The gain of this stage is less than unity, and its output impedance can be designed readily for values as low as 50 ohms. The amplifier, being degenerative, has lower values of input capacitance, is freer from

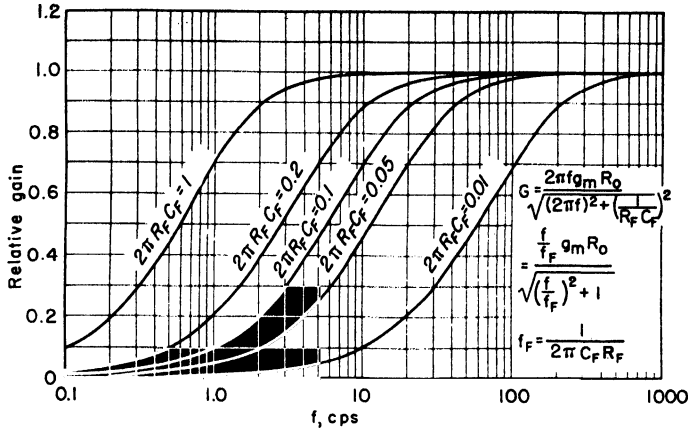



FIG. 34. Amplitude response of l-f compensation system.

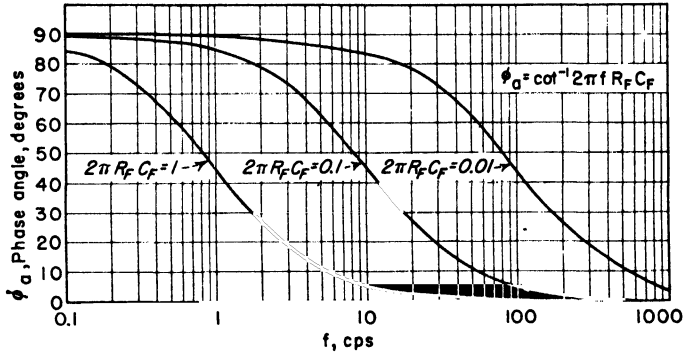



FIG. 35. Phase response of l-f compensation system.

amplitude distortion, and is less affected by changes in supply voltages than is the conventional amplifier stage.

The gain of the cathode-coupled stage is

$$G = \frac{\mu R_k}{r_p + R_k(\mu + 1)} \tag{28}$$

where μ = amplification factor of tube
r_p = its internal plate resistance
R_k = value of cathode resistor

The effective output impedance Z_0 is

$$Z_0 = \frac{R_k r_p/(\mu + 1)}{R_k + r_p/(\mu + 1)} \tag{29}$$

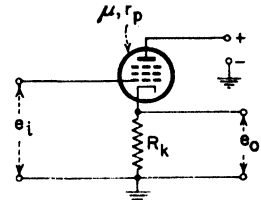



FIG. 36. Fundamental cathode-coupled stage.

An important practical advantage of the cathode-coupled stage is that it may be coupled to the following transducer without the intervention of a coupling capacitor, so that the d-c as well as a-c components of the video signal are transmitted. No pains need be taken to preserve the h-f response, since the low value of impedance makes the shunting effect of the output capacitance negligibly small.

46. Transient Response of Video Amplifiers. The response of a video amplifier to the Heaviside unit pulse of voltage is a general criterion of video-amplifier response.

The response to a single unit pulse is difficult to measure experimentally, but a square wave may be used as the exciting voltage, provided that the period of the wave is long compared with the duration of the transient response. Responses calculated on this assumption are shown in Fig. 37 for a single stage and for several multistage amplifiers. Simple shunt peaking is assumed, for various values of the

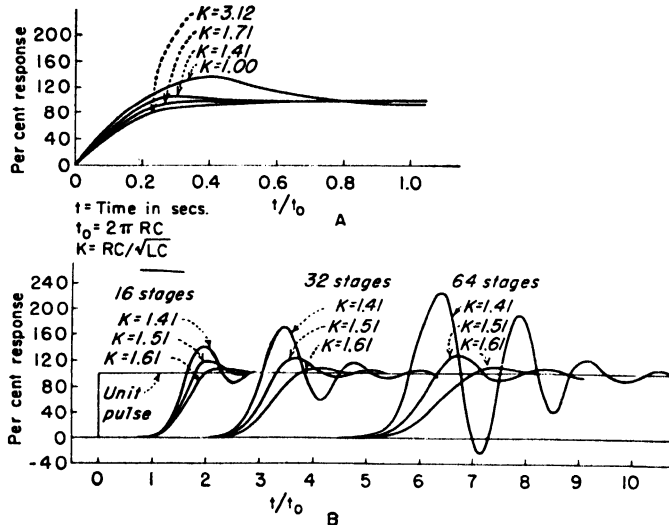



FIG. 37. Transient response of single and multistage compensated video amplifiers.

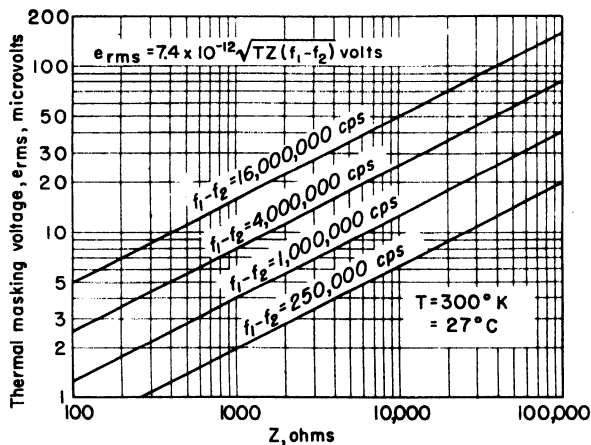



FIG. 38. Thermal agitation voltage generated in wide-band circuits.

parameter $K = RC\sqrt{LC}$, *i.e.*, the ratio of the load resistor R_0 to the impedance of the shunt capacitance C_t at the frequency at which L_0 and C_t are resonant. The case for $K = 1.41$ is equivalent to the cases of $k_R = 1$ and $k_L = 0.5$ (Sec. 41).

47. Noise Limitations to Video Amplification. One of the principal limitations to proper video amplification is inadequate signal/noise ratio. The two sources of cir-

cuit noise, thermal agitation and shot effect, are evaluated in Figs. 38 and 39 for a transmission system responsive to the video range. Values of 50 to 100 μv are common. For a signal/noise ratio of 10:1, commonly assumed as the minimum acceptable for entertainment purposes, the desired signal must accordingly have an rms amplitude of from 0.5 to 1.0 mv.

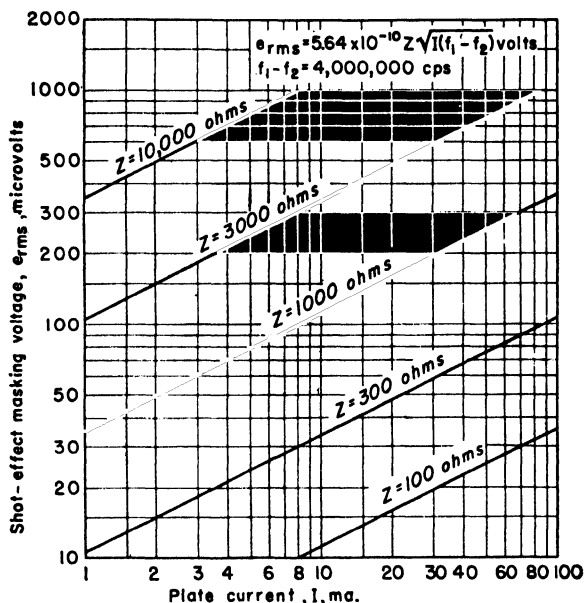



FIG. 39. Shot noise generated in wide-band circuits.

References

Video Amplification:

Barco, A. A.: Measurement of Phase Shift in Television Amplifiers, *RCA Rev.*, **3** (4), 441, April, 1939.

Bedford and Fredendall: Transient Response of Multistage Video Amplifiers, *Proc. IRE*, **27**, 277, April, 1939.

Bereskin, A. B.: Improve High-frequency Compensation for Wideband Amplifiers, *Proc. IRE*, **32**, 608, October, 1944.

Carnahan, C. W.: The Steady-state Response of a Network to a Periodic Driving Force of Arbitrary Shape, and Its Applications to Television Circuits, *Proc. IRE*, **23**, 1393, November, 1935.

Everest, E. A.: Wideband Television Amplifiers, *Electronics*, **11** (1), 16, January, 1938; **11** (5), 24, May, 1938.

Foster and Rankin: Video Output Systems, *RCA Rev.*, **5**, 409, April, 1941.

Herold, E. W.: High-frequency Correction in Resistance-coupled Amplifiers, *Communications*, **18** (8), 11, August, 1938.

Kallman, Spencer, and Singer: Transient Response, *Proc. IRE*, **33**, 169, March, 1945.

Larsen, M. J.: Low-frequency Compensation of Video-frequency Amplifiers, *Proc. IRE*, **33**, 666, October, 1945.

Maclean, W. R.: Ultimate Bandwidths in High-gain Multistage Video Amplifiers, *Proc. IRE*, **32**, 12, January, 1944.

McLachlan, H. W.: Reproduction of Transients by Television Amplifiers, *Wireless Engr.*, **13**, 519, October, 1936.

Schlesinger, K.: Low-frequency Compensation for Amplifiers, *Electronics*, **21**, 103, February, 1948.

Seeley and Kimball: Analysis and Design of Video Amplifiers, *RCA Rev.*, **2** (2), 171, October, 1937; **3** (3), 290, January, 1939.

Swift, G.: Amplifier Testing by Means of Square Waves, *Communications*, **14** (2), 22, February, 1939.

Sziklai and Schroeder: Cathode-coupled Wideband Amplifiers, *Proc. IRE*, **33**, 701, October, 1945.

Wheeler, H. A.: The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *Proc. IRE*, **27**, 359, June, 1939.

———: Wideband Amplifiers for Television, *Proc. IRE*, **27**, 429, July, 1939.

MODULATION, R-F AND I-F AMPLIFICATION, DETECTION

48. Video Modulation. Video modulation is based on the same considerations as audio modulation, with certain specialized requirements. One of the limitations is the small amount of video-signal voltage which may be generated in currently available tubes and circuits. The high capacitance to ground of large water-cooled tubes requires the use of very low values of load resistance to maintain response over the v-f range. The voltage which can be developed across the load resistance depends on the emission current. At present it is difficult to generate more than the 1,500 or 2,000 volts peak to peak over the video range from 30 cps to 4.5 Mc. When high-level modulation is used, therefore, it is usually considered expedient to use grid-circuit modulation, rather than plate-circuit modulation, since the voltage requirements for grid modulation are less by the amplification factor of the modulated stage. Low-level modulation is not similarly restricted but requires a number of linear class B r-f amplifier stages following the modulator.

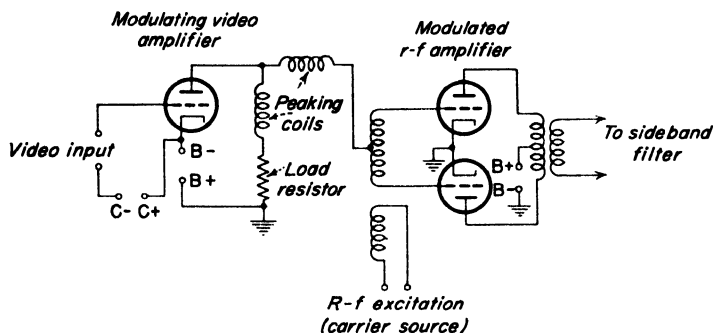



Fig. 40. Fundamental circuit of video modulator.

The second unusual requirement in video modulation is the necessity for maintaining two levels in the modulation envelope at constant amplitudes. These levels are (1) the tips of the sync pulses, *i.e.*, the maximum amplitude of the envelope, and (2) the blanking level or pedestal. Since these levels must remain constant regardless of any changes in the wave form of the camera-signal component, it is necessary to couple the modulating amplifier conductively to the modulated amplifier. This makes necessary a separate power supply for each stage. A typical arrangement is shown in Fig. 40. Here the modulating video amplifier is coupled conductively to the grids of the r-f amplifier. The B supply for the modulating amplifier is in series with the cathode.

At the grid of the modulating amplifier, it is necessary that the blanking level and sync-pulse tip level be constant. The latter levels are caused to assume fixed values by passing the video wave form through a diode rectifier whose cathode is connected to the modulating video-amplifier grid. The load circuit values are chosen so that the rectified d-c potential across the diode assumes a level at the tips of the sync pulses, or just below the tips (the difference being required to supply the diode current). The voltage across the diode forms a part of the fixed bias of the modulating amplifier. The composite wave form, extending more positively than the tips of the sync pulses, causes the modulating amplifier output voltage to extend more negatively than the sync pulses. This output voltage, applied to control the amplitude of the modulated r-f amplifier, causes the sync-pulse tips to assume the peak position in the envelope, while the blanking level and camera-signal components extend to lower levels in the envelope. The sync pulses and blanking level maintain constant amplitudes, whereas the average on the camera-signal component changes with the background illumination of the scene (see Fig. 10).

49. Vestigial-side-band Transmission. The side bands of the modulated r-f signal, assuming a maximum video modulating frequency of 4.5 Mc, extend over a total region of 9 Mc. To conserve space in the ether and at the same time to secure greater efficiency from r-f and i-f amplifiers, the FCC standards specify vestigial-side-band transmission ("single"-side-band transmission). In this system a part of the lower-frequency side band is completely attenuated. By this means the upper side band can be transmitted completely with 4.0 to 4.5 Mc width within the 6-Mc channel assigned by the FCC. A portion of the lower side band, within 1.25 Mc of the carrier frequency, is also transmitted.

The channel composition for vestigial-side-band transmission is shown in Fig. 41, at the top. The lower figure shows the corresponding characteristic of the receiver. In the receiver characteristic the picture signal-carrier voltage is attenuated to 50 per cent of its original level, and the curve of attenuation is linear over a range of 2.5 Mc. This arrangement develops 50 per cent modulation in those portions of the carrier which receive double-side-band treatment (within 1.25 Mc of the carrier). The components further removed from the carrier in the upper side band are inherently 50 percent modulated, so all portions of the signal, when presented to the detector, produce an equal amplitude in the detector output.

FIG. 41. (*Top*) Output characteristic of television transmitter. (*Bottom*) Corresponding response characteristic of receiver.

To form a transmitted signal of the character shown at the top in Fig. 41, a filter having sharp cutoff characteristics is required. The desired upper side band is passed through a capacitor to the antenna, whereas the undesired lower side band is passed through an inductance to an absorbing resistor. Filter structures for this purpose, when employed for high-level modulation, are customarily formed from sections of coaxial transmission lines. In addition to the filter shown, a sharply tuned "notching filter" is usually used to provide additional attenuation at the sound-carrier frequency of the adjacent channel. The characteristic shown in Fig. 41 may also be obtained with low-level modulation by proper positioning of the pass bands of the linear class B stages that follow the modulator.

SEPARATION OF THE SYNCHRONIZING SIGNALS

50. Amplitude Separation. The separation of the composite synchronizing signal from the camera signal is performed after the composite video signal has been developed by the second detector. The composite video signal (Fig. 42*B*) is applied to a "clipper" tube. In Fig. 43 a diode clipper is shown in conjunction with the second detector. It is necessary that the clipping level be maintained continuously at the blanking level to ensure that the camera signal does not affect synchronization and that the maximum amplitude of sync pulses is developed.

51. Wave-form Separation. After the composite synchronizing signal has been separated from the video signal, it is necessary to develop the horizontal sync pulses independently of the vertical sync pulses. The latter separation is carried out by wave-form separation, since the two sets of pulses cannot be distinguished by ampli-

tude means. Essentially wave-form separation depends on circuits which respond to the relative frequency content of the two sets of pulses. The horizontal sync pulses that are of short duration occur 15,750 times per sec and have a predominance of h-f components, whereas the vertical pulses that are of long duration and occur 60 times per sec have a predominance of l-f components. The ratio of the frequencies

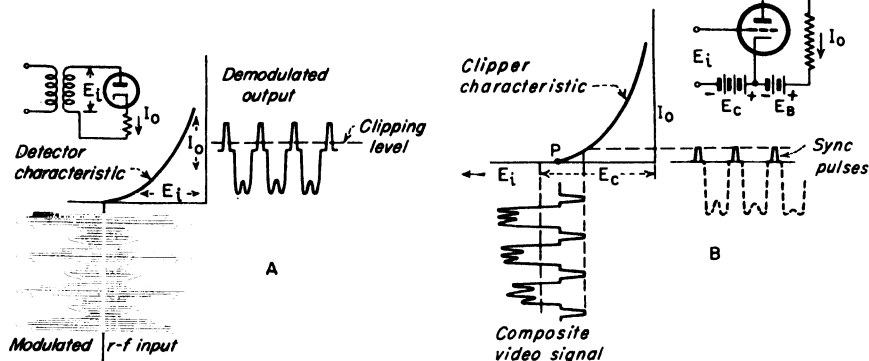



FIG. 42. (*A*) Demodulation of the modulated picture carrier by a diode detector, showing clipping level which separates the camera signal from the synchronizing pulses. (*B*) Clipper circuit and characteristic, used to separate sync signals from camera signals.

of the two sets of pulses 15,750/60 = 262½ is the index of the degree of frequency difference on which the separator circuits may operate.

52. Differentiator Circuit for Horizontal Sync Pulses. The differentiator of Fig. 44 develops the h-f components of the composite synchronizing signal, *i.e.*, the horizontal sync pulses. The series capacitance passes the high frequencies associated with the leading edge of the sync pulse, while retarding all lower-frequency components. The *RC* product (time constant) of the combination is made short compared with the frame-

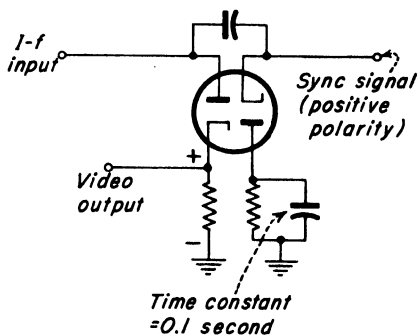



FIG. 43. Combined second detector and sync amplitude separator (clipper).

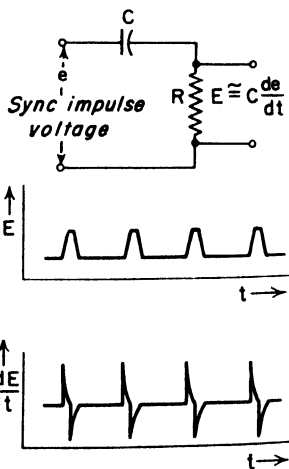



FIG. 44. Differentiator for horizontal sync pulses.

repetition interval (1/60 sec) and long compared with the line scanning interval (1/15,750 sec). The leading edge of the differentiated wave forms is applied, in the proper polarity, to the synchronizing terminal of the horizontal scanning generator.

53. Integrator Circuit for Vertical Sync Pulses. The integrator circuit of Fig. 45 develops a sync pulse from the serrated vertical pulse and equalizing pulses. The

wave forms of input and output are shown. Note that the initial portion of the integrated output pulse is not so sharply rising as that of the differentiated horizontal pulse, and consequently the intersection with the synchronizing control level is not so precisely marked. This makes it necessary to have the wave shape of each successive

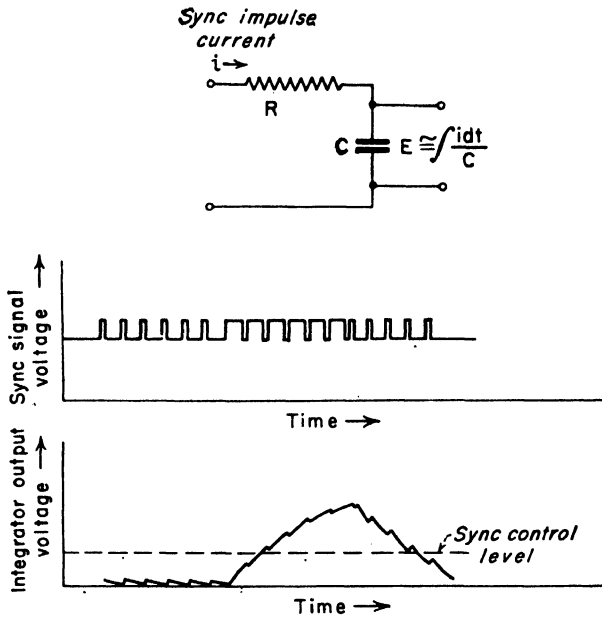



FIG. 45. Integrator circuit and action on vertical sync pulses.

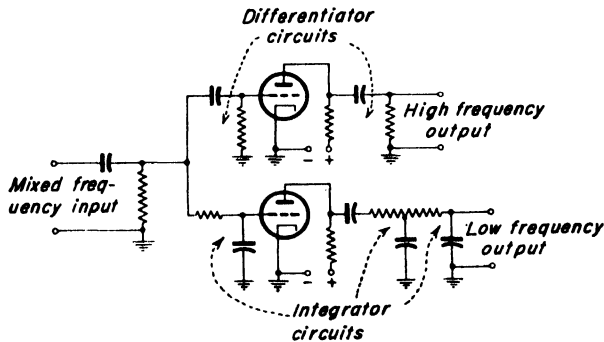



FIG. 46. Combined differentiator and integrator circuits for wave-form separation.

integrated pulse precisely the same. The function of the equalizing in this respect has been pointed out. In addition it is necessary that all traces of the horizontal sync pulses be completely removed from the integrating circuit.

Several differentiating and integrating circuits may be used in cascade to improve the degree of separation. The cascaded circuits may be connected directly together (usually done with integrator circuits), or they may occur in the grid and plate circuits of a sync separator tube as in Fig. 46.

PICTURE TUBES AND ASSOCIATED CIRCUITS

54. Picture Tubes. The conventional picture tube is a funnel-shaped evacuated structure containing an electron gun which forms an electron beam, and a fluorescent screen on which the beam impinges. The beam is deflected by the application of transverse electric or magnetic fields which cause the end of the beam to trace out the interlaced scanning pattern over the fluorescent screen. The current in the beam is capable of variation from zero (cutoff) to a maximum of several hundred microamperes, under the control of the signal potential applied between the cathode and the control electrode of the electron gun.

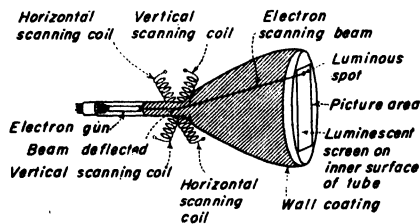



Fig. 47. Structure of typical picture tube.

The beam is deflected synchronously with the scanning agent in the camera tube, and the beam current is controlled by the camera signal. Variations in the beam current produce corresponding variations in the brightness of the fluorescent spot, and the picture is thereby reproduced.

The characteristics of picture tubes depend on the design of the electron gun and on the physical and chemical properties of the fluorescent screen. The electron gun requires a power supply to form the beam. Finally the deflection fields must be provided by scanning generators, and these generators must operate under the control of the synchronizing signals of the video signal.

Picture tubes are classified according to (1) the type of focusing employed (electrostatic or magnetostatic) in the electron gun, (2) the type of deflection (electric or magnetic), (3) the type of phosphor (sulfide and nonsulfide), and (4) the color of the light produced (usually white).

55. Characteristics of Electron Guns. Electrostatically focused electron guns are characterized (1) by the ratio of the voltages applied to the second anode and the first anode. In present designs this ratio varies from 4 to 6. In addition the guns are characterized (2) by the control electrode characteristic which specifies the relation between control electrode voltage and beam current (second-anode current) for different values of second-anode voltage. A typical control characteristic of an electrostatically focused gun employed in the 12AP4 tube is shown in Fig. 48. Curves of this shape are typical of all types of electron guns, whether electrostatically or magnetostatically focused.

Fig. 48. Electron-gun control characteristic; (*A*) 7,000-volt, (*B*) 6,000-volt second-anode voltage.

An important characteristic is the degree of fineness of focus, *i.e.*, the size of the fluorescent spot formed on the screen. Guns of good design are capable of forming a spot about 0.005 in. in diameter, but production tubes usually have spots from 0.01

to 0.015 in. in diameter. The latter spot size permits a picture resolution of 350 lines when the picture height is 6 in. or more (picture width 8 in. or more).

56. Characteristics of Phosphors. The important characteristic of the phosphors (fluorescent materials) employed in picture tubes is the relationship between the

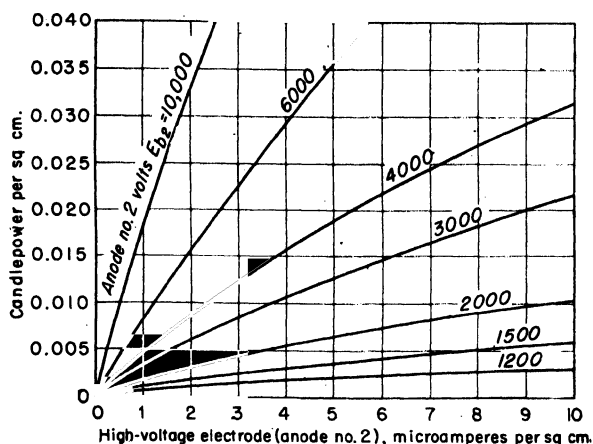



FIG. 49. Type P4 phosphor light-output characteristic.

light produced, the beam current (second-anode current), and the second-anode potential. Figure 49 shows a typical family of such curves, taken for the "P4" white-light phosphor. A recent improvement is the aluminum-backed phosphor, which has superior brightness and contrast when operated at voltages of 9,000 volts or higher.

57. Transfer Characteristic of Picture Tube. The transfer characteristic of a transducer in a television system is the relationship between the significant variational input quantity and the significant variational output quantity. In picture tubes the significant input is the control-electrode voltage, and the significant output is the corresponding light produced on the screen. The transfer characteristic of the 12AP4 tube is shown in Fig. 50. Note that the relationship is not linear but has the "antisaturation" shape. This characteristic tends to enhance the apparent contrast of the picture.

FIG. 50. Typical transfer characteristic of a picture tube, derived from Figs. 48 and 49.

58. Contrast in Picture Tubes. The ratio of the brightness of the brightest portion of an image to the brightness of the darkest portion is called the *brightness-contrast range*. Owing to the effects of light spreading (halation) within the glass envelope the maximum contrast ratio of current tubes is about 100:1. Between closely adjacent portions of the image, halation reduces the maximum obtainable contrast to about 30:1.

59. Dynamic Action of Picture-tube Control Circuit. The dynamic action of the picture-tube circuit is represented by applying the video signal wave form to the transfer characteristic (Fig. 51). The video wave form is applied so that the blanking level corresponds to the zero light (cutoff) point on the transfer characteristic as shown. This bias level must remain fixed. Then the camera signal extending to the right of the blanking level produces light on the screen in accordance with the camera signal, whereas the synchronizing signals to the left of the blanking level are in the infrablack region (beyond cutoff) and do not produce light. The total excursion of the camera signal should be limited so that the control-electrode voltage never becomes positive; usually the control electrode does not go beyond the −5 or −10 volts mark. The average of the picture signal component, taken over the frame interval, establishes the background brightness of the scene, provided the blanking level remains fixed at the light cutoff point.

Fig. 51. Dynamic action of video signal in picture-tube control circuit.

60. D-c Restoration Circuits. Typical circuits used to maintain the blanking level constant at the picture-tube control electrode are shown in Fig. 52. The proportions of C_c and R_g are chosen to have a time constant long compared with the line-scanning interval but short compared with the duration of the changes in background light. In the upper diagram the grid and cathode of the v-f tube act as a diode, whereas in the lower case a separate diode is employed. The diode, acting in conjunction with C_cR_g, develops a bias equal to the peak of the video signal. This peak value remains fixed (assuming no change in signal strength); consequently the remainder of the video signal (the camera-signal components) acts in series with a fixed bias. This fixed bias is coupled conductively (either through the amplifier tube, at the top, or directly, at the bottom) and forms a part of the control-electrode bias. By this means the blanking level remains fixed, and, if the total

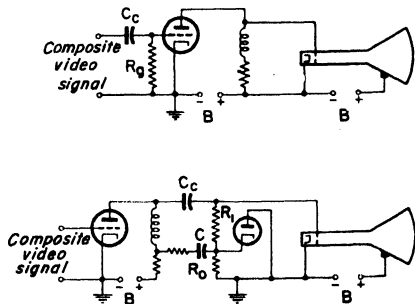



Fig. 52. D-c restoration circuits. (*Top*) Using the video amplifier grid current for rectification. (*Bottom*) Using a separate diode for rectification.

control-electrode bias is fixed so that the blanking level coincides with the light cutoff point, the background brightness of the scene depends only on the average of the camera-signal component, as is required.

61. Picture-tube Power Supplies. The picture-tube power supply consists of (1) a source of high voltage for the first and second anodes, which draw the electrons from the gun and (in the case of electrostatically focused tubes) bring the beam to focus, (2) a source of heater current for the cathode, and (3) a source of focusing coil current (in the case of magnetostatically focused tubes).

A typical high-voltage power supply is shown in Fig. 53. It consists of a single-winding transformer of rms output voltage equal to approximately $V_{dc}/1.4$, where V_{dc} is the desired output d-c voltage; two capacitors of roughly 0.03 to 0.05 μf; a series filter resistor of roughly 100,000 to 500,000 ohms; and a tapped bleeder resistor of about 5 megohms. A resistor in series with the second-anode output tap limits the total output current to a safe

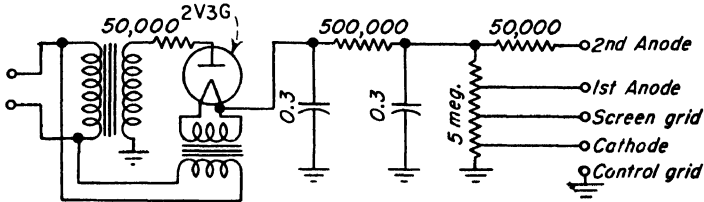



Fig. 53. Typical anode voltage supply for picture tube.

value in case of accidental contact by the operator. The taps required for the various electrodes of an electrostatically focused electron gun are shown.

The current required for the focusing coil of a magnetostatically focused gun depends on the focus-coil design. A typical value is 100 ma at 25 volts, which may be obtained from the receiver proper at the sacrifice of 25 volts in the low-voltage power supply for the receiver.

Heater currents are usually 2.5 volts at 2.1 amp, or 6.3 volts at 0.6 amp.

The high-voltage a-c power may also be derived from (1) high-voltage pulses appearing across the horizontal scanning transformer primary or (2) an h-f oscillator. In the latter case the filtering is simplified, since small values of filter capacitance suffice.

DEFLECTION OF ELECTRON BEAMS

62. Electron-beam Velocity. The amount of deflection suffered by an electron scanning beam depends on the velocity with which the electrons in the beam move. This velocity v is expressed by

$$v = 3 \times 10^{10} \sqrt{1 - \left(\frac{1}{2 \times 10^{-6}E + 1}\right)^2} \quad \text{cm per sec} \qquad (30)$$

where E is the accelerating voltage in volts (approximately equal to the second-anode voltage). This expression takes into account the change in electron mass with velocity. Values of v vary from 2.66×10^9 cm per sec at 2,000 volts to 9.86×10^9 cm per sec at 30,000 volts.

63. Electric Deflection. The deflection d of the scanning beam across the screen of a picture tube caused by passage between parallel deflecting plates is given by

$$d = \frac{1.77 \times 10^{15} E_d l (D + \frac{1}{2}l)}{sv^2} \quad \text{cm} \qquad (31)$$

where E_d = voltage applied to deflection plates, volts
v = electron beam velocity, cm per sec
l = length of deflection plates, cm
s = separation between them, cm
D = distance from screen end of deflection plates to center of screen measured along axis of tube, cm

Typical electrically deflected tubes have deflection sensitivities of from 0.15 to 0.35 mm deflection per volt applied to the deflecting plates, when operated at maximum rated second-anode voltage.

64. Magnetic Deflection. The deflection d across the screen of a picture tube, caused by passage through a uniform magnetic field, is given by

$$d = \frac{1.77 \times 10^7 BlD}{v} \quad \text{cm} \tag{32}$$

where B = flux density of field, gausses
l = its length, cm
D = field-to-screen distance, cm
v = electron-beam velocity, cm per sec

65. Ion Spot. Negative ions liberated from the cathode of the electron gun are focused and deflected in much the same manner as the electrons. In electric deflection the deflection is independent of the charge/mass ratio of the particles; hence the ions and electrons are equally deflected. In magnetic deflection, however, the deflection depends on the square root of the charge/mass ratio. Since the ions have masses several thousand times that of the electron, they suffer correspondingly small deflection. The lack of deflection subjects the center of the scanned area to continual bombardment by the ions, and this eventually results in the formation of a black or yellowish spot. The ion spot is characteristic of the combination of electrostatic focusing and magnetic deflection. Specialized electrode structures have been devised, however, which intercept the ions before they reach the screen. The aluminum-backed phosphor (Sec. 56) is free from ion spot since the aluminum inhibits ionic bombardment.

SCANNING AND SYNCHRONIZATION

66. Saw-tooth Generators. The saw-tooth wave form (Fig. 7) is generated for scanning purposes by the periodic charging and discharging of a capacitor. The charge-time curve is used to produce the active scanning motion, and the discharge curve forms the retrace. To maintain a linear charge curve, it is customary to restrict the charge time to about 0.4 times the RC product of the circuit, or less, and also to make use of the nonlinear dynamic characteristic of the following amplifier to introduce a compensating nonlinearity. Certain forms of multivibrator circuits may be used to produce saw-tooth waves directly.

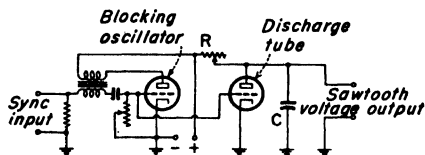



Fig. 54. Blocking-oscillator type of impulse generator.

Usually a separate *discharge tube* is used to discharge the capacitor. The discharge current is passed through a high-vacuum triode whose grid controls the timing of the discharge. The impulses applied to the grid of the discharge tube are usually derived from an impulse generator, although they may consist of the synchronizing signal itself, properly amplified.

Impulse generators used to control the discharge tube in scanning generators take one of two forms, the multivibrator or the blocking oscillator. The blocking oscillator (Fig. 54) consists of a grid-plate-coupled oscillator whose grid is driven negative by the passage of grid current, thus blocking the oscillations suddenly. As the charge leaks off the grid through the grid resistor, the oscillations recommence, to be followed by the sudden blocking of the grid circuit. The sharp impulses appearing between grid and ground are used to control the discharge tube as shown.

Fig. 55. Multivibrator type of impulse generator.

A multivibrator type of saw-tooth generator is shown in Fig. 55. This circuit oper-

ates by virtue of the connection between the plate circuit of the output tube and the grid circuit of the input tube. The alternate charge and discharge of the coupling capacitor can be used to produce either impulses or saw-tooth waves, depending on the circuit constants.

67. Production of Current Saw-tooth Waves. Saw-tooth waves of voltage produced by saw-tooth generators suffice to deflect the beam of an electrically deflected tube, which is a voltage-operated device, provided only that the peak-to-peak value of the saw-tooth wave is great enough to produce full deflection. In magnetically deflected tubes the deflection is proportional to the current in the deflection coils; hence saw-tooth waves of current are required. The voltage wave form required to produce saw-tooth waves of current depends on the inductance and resistance present

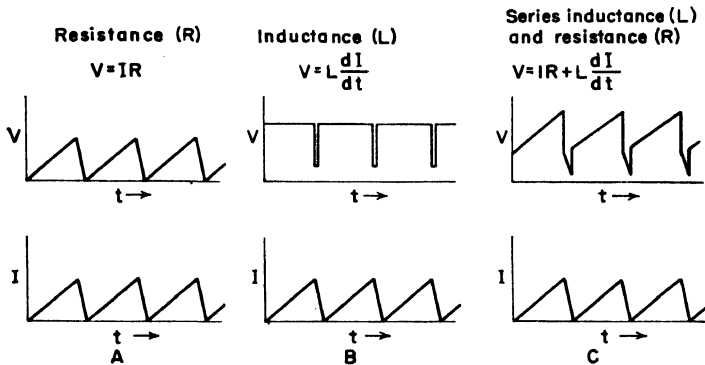



FIG. 56. Saw-tooth waves of current and corresponding voltage wave forms in circuits of *R*, *L*, and *LR*.

in the scanning-coil windings. An "impulse" voltage wave is required for coils exhibiting a large inductance-resistance ratio. For lower L/R ratios the voltage wave form is a combination of impulse and saw-tooth waves. The several voltage and current wave forms for these cases are shown in Fig. 56.

The part-impulse part-saw-tooth wave form may be produced simply by applying a saw-tooth wave to a series *RC* combination. The saw-tooth component develops across the resistance, while the impulse portion develops across the capacitance.

The presence of distributed capacitance in the scanning-coil windings gives rise to resonance oscillations when the impulse voltage wave form is applied. These residual oscillations may be damped out by connecting a rectifier tube and a shunt *RC* circuit in series across the scanning-coil terminals.

68. Amplification of Scanning Wave Forms. The preservation of the scanning wave form in the amplifier subsequent to the scanning generator is based on the considerations for video amplifiers. Usually it is desirable to pass the fundamental and 20 harmonics, which makes the range 60 to 1,200 cps for the vertical scanning amplifier and 15,750 to 315,000 cps for the horizontal amplifier. The phase and amplitude characteristics must be linear over these ranges.

For electric deflection it is essential that the scanning-generator output be disposed symmetrically with respect to the deflection plates, and this is carried out by push-pull amplification. The center point of the push-pull output is connected through a high resistance to the second-anode terminal of the picture tube. Care must be taken to allow the necessary peak-to-peak voltage to develop across the amplifier output without breakdown of insulation and excessive stress in the tube structures. The necessity for high scanning voltages has limited application of electric deflection to tubes operating below 6,000 volts, second-anode voltage.

In magnetic deflection, heavy current rather than high voltage is required to secure full deflection. To secure the current, it is customary to employ a voltage step-down

transformer in the output of the scanning amplifier. This transformer must meet the amplitude- and phase-frequency characteristics of the amplifier itself. High voltage develops across the primary of this transformer as a result of the rapid changes of current in the secondary. The amplifier tubes and other components must be capable of withstanding these voltage peaks, which often attain several thousand volts amplitude.

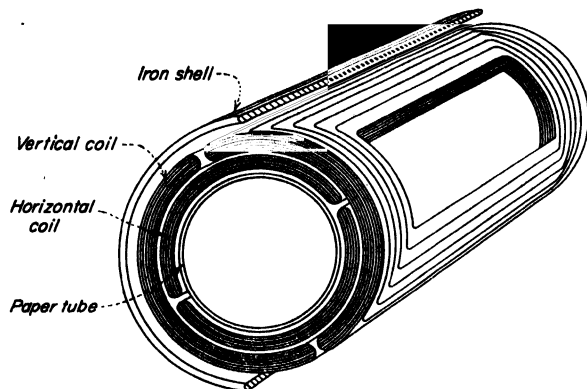



Fig. 57. Construction of typical scanning yoke (magnetic-deflection coil system).

69. Scanning Yokes. The set of coils required for magnetic deflection, called a *scanning yoke*, consists of two sets of coils. One, arranged about a vertical axis transverse to the tube axis, produces the horizontal deflection, and another, arranged on a horizontal axis transversely to the tube axis, produces the vertical deflection.

Among the factors on which yoke design depends are (1) the angle of deflection required, which determines the required number of ampere-turns as well as the allowable physical length of the yoke; (2) the necessity of providing a uniform field, to avoid defocusing the spot and distorting the orthogonal shape of the scanning pattern; and (3) the proportioning of the L/R ratio to secure linear deflection with a given deflection amplifier and output transformer.

CONTRAST AND GRADATION OF TELEVISION IMAGES

70. Over-all Brightness Transfer Characteristic. The ability of the television system to reproduce brightness contrasts and tonal gradations is expressed by the over-all brightness transfer characteristic (Fig. 58). The ordinates give the range of bright-

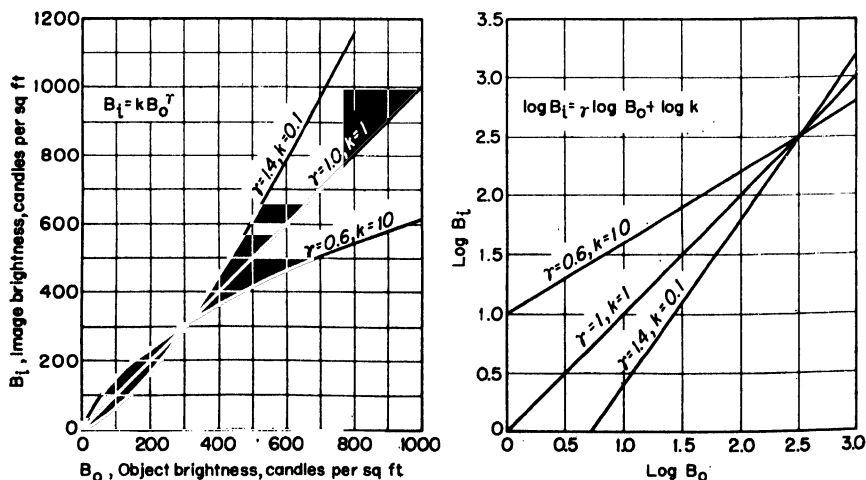



Fig. 58. Brightness-transfer characteristics of a television system.

ness in the reproduced image (image brightness) corresponding to the range of brightness in the original object (object brightness) plotted in the abscissas.

The actual shape of this curve depends on the transfer characteristic (input-output relationship) of each item in the system. In general the actual characteristics cannot be expressed in simple analytic form. However, if an idealization is made, the curves may be expressed in the following form:

$$B_i = B_0 k_0 \gamma_0 \quad (33)$$

where B_i is the image brightness corresponding to the object brightness B_0, k_0 is the proportionality factor relating the image brightness scale to the object brightness scale, and the

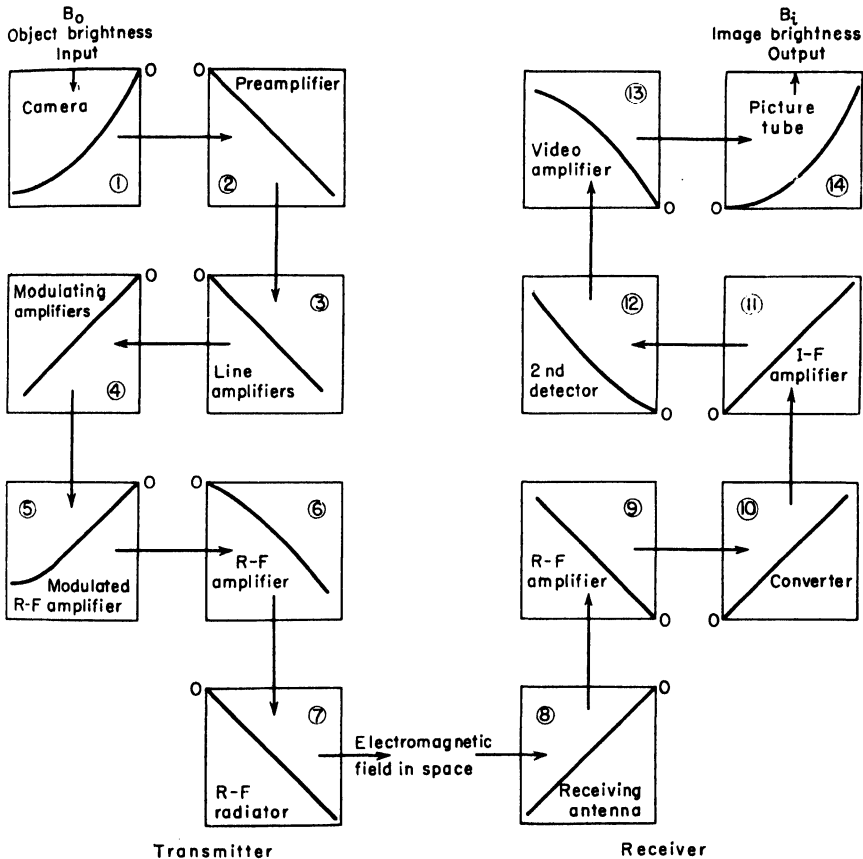



FIG. 59. Transfer of object brightness to image brightness through subsidiary transfer characteristics of the elements of a television system.

exponent γ_0 ("transfer slope") determines the extent and direction of the curvature of the characteristics. For unity slope ($\gamma_0 = 1$) the relationship between B_i and B_0 is linear. For slope greater than unity, the curve has an "antisaturation" shape; for slope values below unity, the curve has a "saturation" shape.

The value of slope determines the subjective contrast of the image as viewed by the observer, since the sensation of light in the mind is approximately proportional to the logarithm of the brightness. When Eq. (33) is expressed in logarithmic form

$$\log B_i = \log k_0 + \gamma_0 \log B_0 \quad (34)$$

all the relationships between log B_i and log B_0 become linear and the slope of the lines is directly proportional to the slope value of γ_0. In consequence high contrast is produced by correspondingly high values of slope.

71. Subsidiary Transfer Characteristics. The input-output characteristics of each subsidiary item of equipment in the system can be expressed by a similar relationship:

$$\text{Output} = k(\text{input})^{\gamma}$$

where k relates the scales of the input and output quantities and γ is the slope exponent describing the curvature of the characteristics. By combining each curve in the transmission system successively, equating the output of one device to the input of the succeeding device, it can be shown that the over-all slope of the system is equal to the product of all the subsidiary slopes. In consequence, the effect of one item of equipment whose slope is lower than unity may be compensated by that of another whose slope is the inverse of the first. The slope of iconoscope tubes, for example, lies at about 0.7, whereas that of picture tubes is about 2.5. Assuming that the subsidiary amplifiers, modulators, and demodulators are linear (slope unity), the over-all slope is then $0.7 \times 2.5 = 1.8$; *i.e.*, the slope is somewhat above unity. The orthicon camera, on the other hand, has a slope of unity, and the over-all slope in this case would be 2.5, producing a considerably more contrasty image. The desirable value of over-all slope, following motion-picture practice, is between 1.2 and 1.7. The high value of slope aids in restoring color contrasts lost through the monochromatic nature of the reproduction. It should be noted that high contrasts are limited by picture-tube performance.

The values of the subsidiary slopes also bear on the signal/noise ratio of the system. If a transmitter slope less than unity is employed, most of the picture information consists of signal excursions having amplitudes high on the dynamic characteristic, above the noise. A compensating higher value of slope in the receiver may be used to produce an over-all value within the desirable range of 1.2 to 1.7.

References

Picture Tubes and Image Reproduction:

Bachman and Carnahan: Negative Ion Components in the Cathode Ray Beam, *Proc. IRE*, **26**, 529 May, 1938.

Brewer, W. L., and J. H. Ladd: Comment on NTSC Specifications for Color TV, *Proc. IRE*, January, 1955.

Burnett, C. E.: A Circuit for Studying Kinescope Resolution, *Proc. IRE*, **24**, 1095, August, 1937.

Epstein, D. W., P. Kaus, and D. D. Van Ormer: Improvement in Color Kinescope through Optical Analogy, *RCA Rev.*, December, 1955.

Hurford, W. L., and R. J. Marian: Monochrome Vidicon Film Camera, *RCA Rev.*, September, 1954.

Iams, H.: A Fixed Focus Electron Gun for Cathode-ray Tubes, *Proc. IRE*, **27**, 103, February, 1939.

Kozanowski, H. N.: Vidicon Film-reproduction Cameras, *J. SMPTE*, February, 1954.

———: Three Vidicon Color Film Camera, *Broadcast News*, May–June, 1954.

Langmuir, D. B.: Theoretical Limitations of Cathode-ray Tubes, *Proc. IRE*, **25**, 977, August, 1937.

Leverenz, H. W.: Problems Concerning the Production of Cathode-ray Screens, *J. Opt. Soc. Am.*, **27**, 25, January, 1937.

Levy and West: Luminescence and Its Application to Television, *J. Telev. Soc.*, **2**, 337, March, 1938.

MacAdam, D. L.: Color Balance for TV, *Proc. IRE*, January, 1955.

Maloff, I. G.: Gamma and Range in Television, *RCA Rev.*, **3** (4), 409, April, 1939.

——— and D. W. Epstein: Theory of the Electron Gun, *Proc. IRE*, **22**, 1386, December, 1934.

——— and ———: "Electron Optics in Television," Chap. XII, McGraw-Hill, 1938.

Neuhauser, R. G.: Vidicon for Film Pickup, *J. SMPTE*, February, 1954.

———: Optimum Performance of Image Orthicons and Vidicons in Broadcast Service, *RCA Rev.*, May, 1954.

Torre, A. J., and L. R. Kirkwood: The CT-100 Commercial Color TV Receiver, *RCA Rev.*, September, 1954.

Zworykin and Painter: Development of the Projection Kinescope, *Proc. IRE*, **25**, 937, August, 1937.

COLOR TELEVISION

72. Compatibility with Monochrome System. The color television standards in use in the United States provide for compatible color and monochrome transmissions. Compatibility is achieved by processing the color signal in such a way that it contains one component which conveys luminance information plus two other components which convey color-difference or chrominance information. The luminance component is very similar to a conventional monochrome signal, and it is transmitted in

such a way that it is capable of producing high-quality images on conventional monochrome receivers. The color-difference or chrominance components are transmitted in the form of side bands around a so-called color subcarrier of 3.579545 Mc. These side bands are simply added to the luminance component of the signal, but the possibility of interference effect is minimized by the use of a technique known as frequency interlace. The chrominance components are ignored by monochrome receivers but are utilized in color receivers, through suitable filters and demodulators, for the generation of signals suitable for controlling red, green, and blue elements in a tricolor display. In the absence of the chrominance components (implying a monochrome transmission), a compatible color receiver automatically produces a monochrome image.

The sections which follow describe the requirement for color-television transmission and reproduction and the means by which these requirements are met.

73. Color Matching. In color television, the system must be able to achieve a match between colors in the original subject and in the televised reproduction. Color matching is greatly simplified by the fact that color perception in the human eye is based on three sets of color receptors in the central portion of the retina. These are sensitive over three contiguous portions of the visible spectrum, centering in the red, green, and blue regions. The spectral sensitivity of each group of receptors may be measured by noting the relative amounts of three spectral primary colors required to match the various colors of the spectrum. Figure 60 shows the relative amounts of standard red, green, and blue spectral primaries required to match each color in the spectrum, based on the "standard observer" of the International Commission on Illumination (ICI).

By so combining three primary colors it is possible to produce a very close match to any color in the spectrum, as well as colors (such as brown, purple, magenta, cyan) which do not occur in the spectrum.

The range of colors over which such matches can be made depends on the number and character of the primary colors used. Two primary colors suffice to match only a very narrow range of colors, while three properly chosen primaries suffice to match closely nearly all the colors that occur in nature. More than three primary colors serve no practical purpose.

A convenient graphical representation of the color-matching process is the color diagram (properly, "chromaticity diagram") shown in Fig. 61. This is a two-dimensional representation of three quantities, x, y, and z, known as the trichromatic coefficients. The quantities x and y are shown explicitly on the diagram, and the z coordinate is given implicitly by the relation $x + y + z = 1$. These quantities specify in normalized fashion the relative proportions of three standard primaries required to match particular colors. A point on the diagram represents one such mixture of primaries and corresponds to the color matched by that mixture. If points are plotted, representing the mixtures required to match each color as it occurs in the spectrum, the points fall on a locus ("spectral locus") shown as a solid line in the figure. The blue colors occur near the origin of the coordinates, the green colors at the top, near the y axis, and the red to the right in the diagram. The spectral locus is closed by a straight line between its extremes. All colors visible to the ICI standard observer may be represented as points in the area bounded by this roughly triangular figure.

The convenience of the color triangle resides in the fact that a straight line drawn between points representing two colors contains points representative of all the colors which can be obtained by combining the two end-point colors in various proportions. Moreover, if three points are chosen and lines are drawn between them to form a triangle, the triangle so formed encloses points representative of all the colors that can be formed by mixing the three apex-point colors in various proportions. In particular, if lines are drawn from an interior point to the apex points, the lengths of these lines are in inverse ratio to the amounts of each apex-point color required in the mixture to match the color represented by the interior point.

The standard ICI spectral primaries (on which Fig. 60 is based) are shown in Fig.

61 as B_s (blue, 435.8 mμ), G_s (green, 546.1 mμ), and R_s (red, 700 mμ). The triangle formed by these points contains all the colors that can be matched by simple addition of these primaries. Although a portion of the area within the spectral locus is not included in the triangle, the portion not included contains shades of green and blue which are not clearly distinguished from one another by the eye, and which in any event do not occur in nature to any extent. The colors included within the three primaries include the vast majority of distinguishable hues and saturations; omission of colors external to the triangle formed by the primaries does not preclude accurate reproduction of colors encountered in practice.

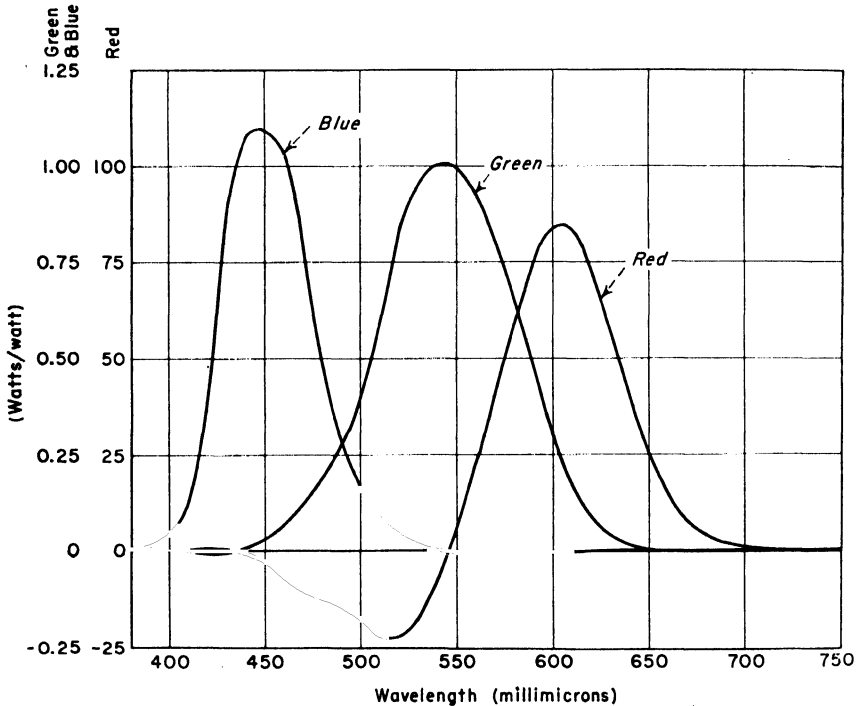



FIG. 60. Relative amounts of standard primary colors required to match spectral colors as per ICI "standard observer."

White light is represented in the central region of the chart. The point marked C represents a standard white (ICI "illuminant C") which closely approximates the color of daylight from the northern sky plus sunlight. Any three colors surrounding this point will, when combined in the proper proportions, match this white. In fact any two colors, the line between which passes through C, will, when properly combined, match this standard white. But such a two-primary system will not match colors off the line and so gives a very poor approximation to nature. For this reason the two-primary system, although occasionally used for motion-picture work, was abandoned in color-television development.

In color photography and television it is not convenient to employ spectral primary colors, which can be obtained only by selecting a narrow portion of a spectrum produced by a prism or grating. Rather, primary colors are produced by passing light through filters, *i.e.*, transparent pieces of colored gelatin, glass, or plastic. These

filters are interposed between the subject and the camera. Standard filters of this type, such as the Wratten series, are widely available.

In a television system, the primary colors used at the transmitter are compounded of three characteristics: the spectral output of the light source, the spectral transmission characteristic of the filter material, and the spectral sensitivity of the photosensitive surface in the camera. At the receiver the corresponding primaries are compounded of the spectral output of the cathode-ray phosphor. The area contained

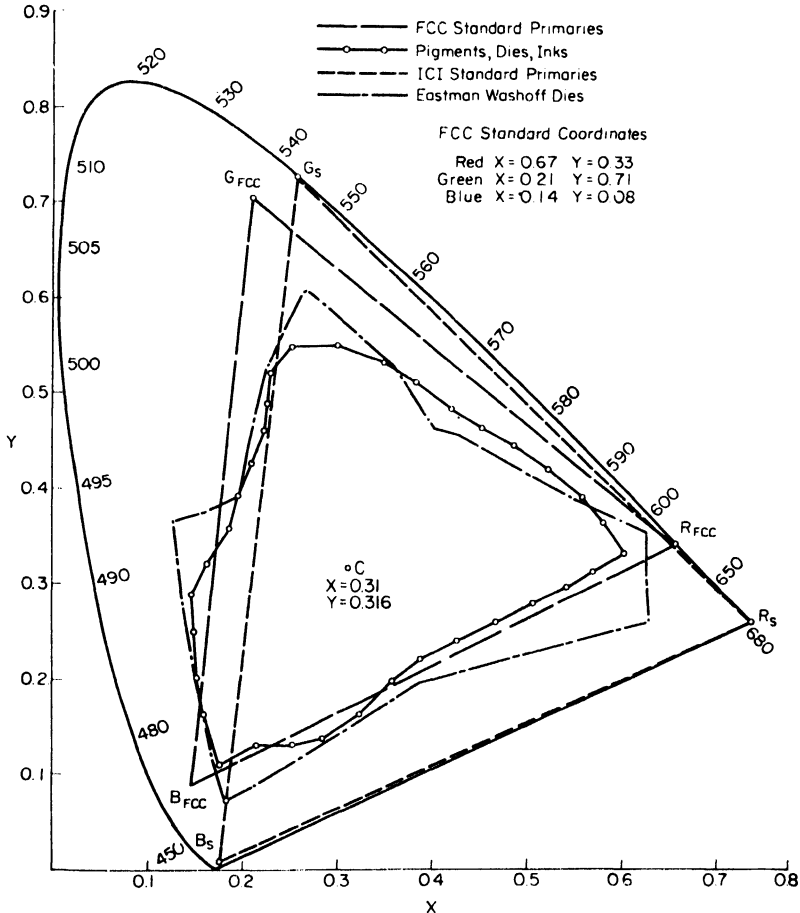



FIG. 61. Chromaticity diagram showing color gamuts.

by the FCC television primaries is greater than that covered by the photographic transparency processes (Technicolor, Kodachrome, Ektachrome, or Anscochrome). The superiority of the color gamut of the television system over printing inks, pigments, etc., is readily apparent in Fig. 61.

74. Color Balance. In accordance with modern television practice, *luminance* is a measure of the light flux emitted, reflected, or transmitted per unit solid angle per unit projected area of the source. The usual units are lamberts, foot-lamberts, or candles per square foot. This quantity is also called *photometric brightness.*

Brightness is defined as the attribute of visual perception in accordance with which an area appears to emit more or less light.

The use of these terms permits "brightness" to be used entirely with reference to the sensory response. A photoelectric instrument calibrated in foot-lamberts would not be referred to as a "brightness" meter. If correctly calibrated it is a "luminance" meter.

In color reproduction it is necessary that the color match between an original subject and the televised reproduction be correct over the range of luminance present in the original scene, and also that the colors appear relatively correct when the luminance range of the reproduction is substantially less than that of the original scene. The color-matching process represented in the color diagram applies over a wide range of absolute luminance. As an example, consider that a given blue is matched by a specific mixture of primaries. If the luminance of the blue is substantially increased so that it appears to be blue-white, the same mixture of primaries will match the blue-white, provided only that the luminance of each primary is correspondingly increased in proportion.

To achieve color balance over a range of luminance the intensity of the primaries must vary in proportion to the luminance of the original color. Thus, if the transfer characteristic that applies to the green primary does not have the same extent and shape as those applying to the red and blue, correct color matching would be achieved only over a restricted range of luminance. To avoid this restriction, dynamic-control amplifiers may be employed to match the shape and extent of all the transfer characteristics. To permit accurate delineation of colors over a range of luminance, it is required that all three transfer characteristics be linear, *i.e.*, the log of each receiver primary luminance should be a linear function of the log of the corresponding transmitter primary luminance.

A critical test of the ability of a color-reproduction system to maintain color balance is the transmission of a gray scale, *i.e.*, a series of gray patches having steps of luminance in logarithmic progression. Lack of color balance is revealed by the presence of color in one or more of the patches of the reproduced scale. Adjustment of the transfer-characteristic control amplifiers to remove any color tinge from the steps of the gray scale is then indicated.

75. The Color Subcarrier Frequency. For reasons which will be explained, the FCC color-television standards specify a horizontal scanning rate of 15,734.264 cps and a vertical field rate of 59.94 cps, instead of the monochrome rates of 15,750 and 60 cps.

The color subcarrier frequency of 3.579545 Mc is spaced less than 4 Mc from the video carrier. If the color subcarrier frequency were slowly changed, it would be observed on a picture monitor that the beat-note interference would vary widely in visibility. The subcarrier frequencies at which the beat pattern is minimized are odd multiples of half the horizontal line frequency. One half of the color line frequency of 15,734.264 is 7,867.132 cps. The 455th multiple of this frequency is 3.579545 Mc. During each horizontal line there are 227.5 cycles of video carrier-color subcarrier beat. Since each cycle represents two changes of light intensity, there are 455 variations of brightness per line.

Upon the completion of one field scan, the phase of the beat note reverses 180 deg because the color subcarrier is an odd multiple of half the line frequency. As a result the lines of the following field contain a beat-note pattern having the amplitude variations reversed. Since the persistence of vision of the human eye is long compared with these brightness changes, they blend and the eye sees only the average fine-detail brightness. To all intents the beat pattern is invisible. Since there are an odd number of beat-frequency light variations per line, they are staggered on successive lines. This provides added cancellation.

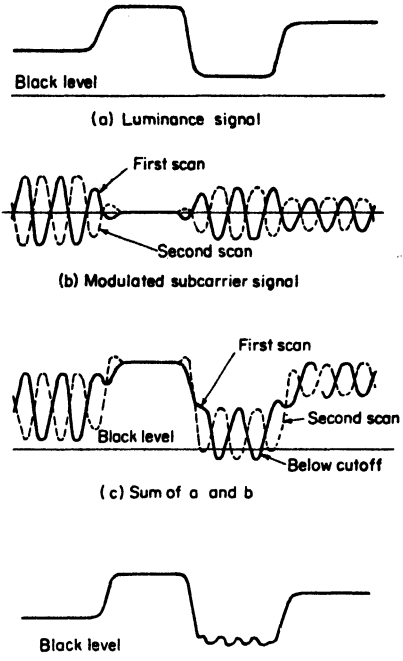
Another possible source of beat-note interference is present between the color subcarrier and the sound carrier. The difference in frequencies is only slightly over 900 kc and the interference produced would be very objectionable. This may also

be minimized by frequency interlace but to do so and also maintain frequency interlace with the video carrier frequency it is necessary to reduce the field rate from 60 to 59.94 and the line rate to 15,734.264 as stated at the beginning of this section. These small changes in transmission standards between monochrome and color transmission and reception are well within the system tolerances and introduce no problems. Figure 62 illustrates the manner in which frequency interlacing produces beat-interference cancellation.

76. Suppressed Subcarrier Transmission. The United States color standards employ suppressed subcarrier transmission through the use of color subcarrier balanced modulators. Since only the color side-band energy is transmitted, no subcarrier signal is present when the camera scans a black, white, or gray object. Also, when scenes of low color saturation are scanned by the color camera, the energy in the color side bands is correspondingly low. This feature aids materially in the suppression of beats between the color signal and the sound signal because the subcarrier signal is present only when color-difference information is actually being transmitted, and is of low amplitude for colors of low or intermediate saturation.

FIG. 62. Manner in which beat interference is minimized by frequency interlacing.

77. Color Information Transmitted. Highly satisfactory color reproduction may be accomplished through proper mixtures of the red, green, and blue primary colors. Good color reproduction also could be obtained if other primary colors were adopted so long as no one of the primaries could be produced by any combination of the other two. The luminance of a picture consists of the integrated luminances of the three colors of which it is composed. In the United States color-television system the receiver picture tube is fed by a luminance component accompanied by color components of red, green, and blue. However, the signal transmitted, prior to the receiver matrix, consists of only a luminance and two color-difference (chrominance) components. The color components may be eliminated to leave only the luminance achromatic component, which was the only form of television picture transmitted and received prior to the advent of color television. A black-and-white television camera scanned colored images but selected only the luminance component, rejecting the color components.

The reproduction of a colored scene on a monochrome reproducing system requires, for most natural rendition, that the proper luminance values of each of the three primary colors be transmitted and reproduced. In a color system, substantially equal parts of red, green, and blue are arbitrarily equated to the amount needed to match a specified white. Because of the characteristics of our eyes and brain, red, green, and blue colors of corresponding luminance values do not appear to us to be equally bright. A monochrome system sees everything as black or white or a shade of gray. Viewers in their homes do not see the original object but only a reproduction of it. The hue and saturation of costumes, scenery, and background objects may vary widely at the discretion of the television producer. Considerable latitude is possible in creating the scene desired. Therefore, the precise reproduction of the luminance values of such objects is not highly important. A desirable range in

brightness contrast between objects is more important to separate the parts of the scene being viewed. However, in transmitting flesh tones, large inaccuracies in reproducing the luminance values of red, green, and blue may give an unnatural emphasis to facial features which is noticeable to viewers familiar with relatively uniform facial characteristics.

Equal luminance values of saturated red, green, and blue colors produce a brightness sensation to the human seeing mechanism of 59, 30, and 11 per cent, respectively. The use of these values in the United States television system for the luminance components of color pictures produces a natural and satisfactory reproduction on monochrome receivers. These values were chosen to facilitate the use of the constant-luminance principle. The choice of these coefficients for the monochrome signal facilitates the use of gains in the receiver which minimize the effects of sine-wave interference in the color channels. Horizontal bars caused by such interference would have color values in a color receiver but the brightness value would be minimized. Similarly, the visibility of high-frequency thermal noise or high-frequency components of the brightness signal which enter the receiver decoders are minimized. This follows from the constant-luminance principle upon which the United States system is based. In this system signals in the color subcarrier channel have very little effect upon the luminance of the reproduced picture. The luminance is almost completely controlled by the picture carrier channel, not the color subcarrier channel.

The constant-luminance principle takes advantage of the fact that the eye is much more sensitive to luminance variations than to color variations. The adjustment of the luminance coefficients in the system is accomplished by control of the gain of the separate amplifiers which amplify the outputs of the red, green, and blue cameras.

The luminance component of a color picture is transmitted independently by modulation of the picture carrier. The color components are transmitted as color-difference signals by modulation of the color subcarrier. A color-difference signal is one which, when added to the monochrome signal, produces a signal representative of one of the tristimulus values (with respect to a stated set of primaries) of the transmitted color. A blue color-difference signal voltage may be obtained from the blue camera tube by subtracting from it the voltage of the luminance signal. Similar operations may be performed at the output of the red and green camera tubes to obtain the red and the green color-difference signals. However, it is not necessary to transmit the separate red, green, and blue components to reproduce tricolor pictures. It is only necessary to transmit the luminance component and two color-difference subcarrier components. This technique is employed. By limiting the amount of information transmitted to only the essential ingredients, which in the United States system are the luminance and two color-difference components, the system attains maximum simplicity and spectrum economy. The two color components are transmitted by modulation of the color subcarrier in a manner to be described.

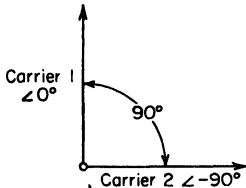



FIG. 63. Two-phase modulating system employing separate amplitude-modulated carriers displaced 90 deg. This is equivalent to having simultaneous amplitude and phase modulation of a single carrier.

78. Modulation of Color Subcarrier. The luminance component is transmitted by amplitude modulation of the video carrier. Two color components must also be transmitted. It is not possible to employ two separate modulated color subcarriers of different frequency satisfactorily because there would be a difference of frequency between them which would be an even multiple of one-half the line frequency. This would not be frequency-interlaced and, therefore, the interference beat-note pattern would not be self-canceling. Any nonlinearity in the system such as exists, for example, in picture tubes would produce this beat note representing the difference frequency between the two subcarriers. This difficulty is avoided by the use of two-

phase modulation of a single subcarrier. This two-phase modulation is accomplished by using two subcarriers of identical frequency but displaced in phase by 90 deg as shown in Fig. 63. In practice the two subcarriers are produced from a single source with two branched outputs. On one the phase is unchanged and on the other it is shifted 90 deg. Figure 64 is a block diagram of the modulating system. The output of this two-phase modulating system is the vector sum of the two quadrature-related and separately amplitude-modulated subcarriers of the same frequency. In the resulting vector sum, the amplitude of the modulation products is a function of color saturation while the resulting phase represents hue. At the receiver the two color components may again be separated and re-created by synchronous demodulators. The in-phase a-m signal is demodulated by applying to the synchronous detector an unmodulated 3.579545-Mc subcarrier of the same phase as was used in the original amplitude modulation. The quadrature-phased modulated signal is demodulated in the same manner except that the subcarrier is shifted 90 deg to correspond exactly to the phase in the original modulation. This imposes the requirement that there must be available at both transmitter and receiver subcarrier voltage of identical frequency and phase. It is impractical to attain such precise control of unmodulated subcarrier frequency and phase by the use of a free-running subcarrier oscillator at the receiver. Therefore, a means is required to control precisely a receiver subcarrier oscillator from the original 3.579545-Mc transmitter oscillator. This is accomplished by transmitting bursts of unmodulated subcarrier power from the transmitter to the receiver where it locks the frequency and phase of the receiver oscillator.

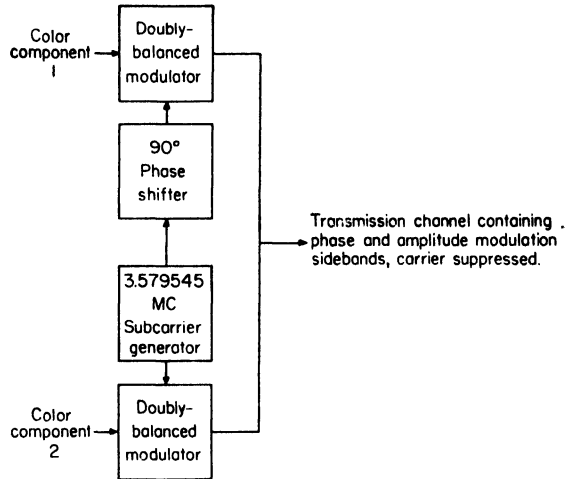



Fig. 64. Block diagram of two-phase subcarrier modulating system.

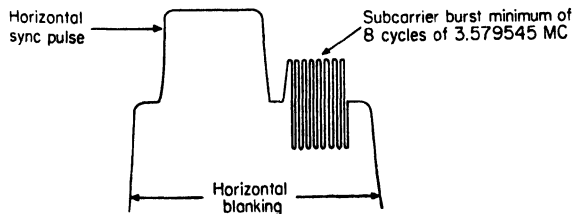



Fig. 65. Subcarrier burst as transmitted on horizontal blanking pedestal.

79. Subcarrier Burst. The subcarrier bursts are transmitted by modulation of the horizontal blanking pedestal as shown in Fig. 65. From 8 to 11 cycles are transmitted. The burst voltage is suppressed at other times. The insertion time and interval are controlled by an electronic gate which is in turn keyed open by burst flag pulses derived from the horizontal deflection circuits. At the receiver a similarly controlled pulse gate isolates the subcarrier bursts from all other voltages at times when the burst is being transmitted and received. By example, a block diagram of a demodulating system is shown in Fig. 66. It may be noted that the receiver subcarrier oscillator phase is locked to the burst phase by means of a phase detector and reactance tube.

In these circuits the burst and the oscillator phases are compared in the phase detector. Any difference produces a succession of pulses which occur at the gating frequency. These are filtered to produce a d-c control voltage which causes the reactance tube to correct the oscillator phase within a few degrees.

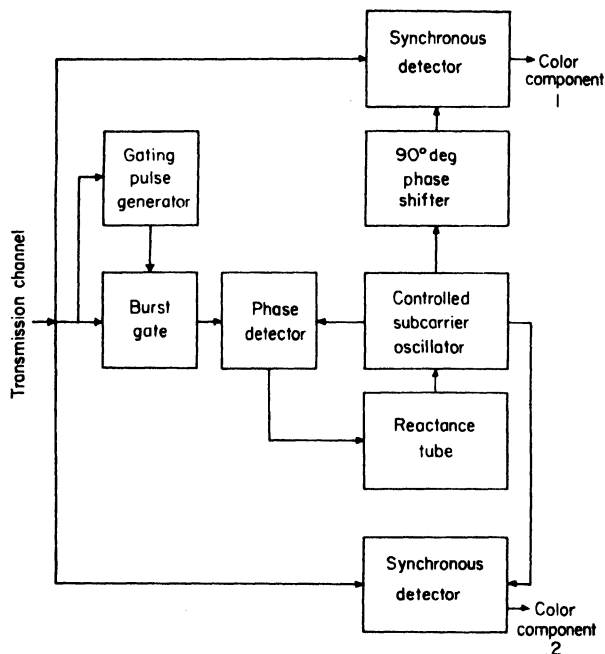



FIG. 66. Block diagram of a two-phase subcarrier demodulating system.

80. Synchronous Detectors. When suppressed carrier transmission is employed and only the modulation side bands are transmitted, the original modulating signal may be re-created only if a locally generated carrier is added at the receiver detector. The carrier added must be of the same frequency or synchronous with the suppressed carrier, and, for full utilization of both side bands, the phases must also correspond. This type of detection, in which a synchronous locally generated carrier is added to incoming side-band energy, is called synchronous detection. If the locally generated carrier is in quadrature with the suppressed carrier, the upper- and lower-frequency side bands do not add. They are in effect shifted +90 deg in one case and −90 deg in the other and cancel, producing no output. The two-phase modulating system described above depends upon this characteristic. The side bands of color component No. 1 produce normal demodulated output when combined with a locally generated carrier corresponding in phase to the original carrier, but produce no output when combined with a carrier rotated 90 deg, such as that of color component No. 2. The reverse is true. The United States standards of color television permit the use of this principle. If the two sets of side bands correspond in band width and both sets of side bands are transmitted normally for both color components, no cross talk between channels occurs. However, if on one channel the upper or lower side bands were missing or attenuated, complete cancellation would not occur when combined with the phase-shifted carrier of the other channel, and cross talk would then occur. In the United States system, one color channel employs vestigial-side-band transmission by attenuation of the highest-frequency components. However, in this

case cross talk is prevented by band-limiting the opposite channel so that the attenuated side-band frequencies are prevented from reaching the synchronous detector.

81. *I* and *Q* Signals. The in-phase color component which is transmitted is referred to as the *I signal* and the quadrature component as the *Q signal*. The brightness component is referred to as the *Y signal*. The outputs of the red, green, and blue cameras are fed to a color-processing matrix in which the three voltages are combined in proper proportions to produce the *Y* signal and also the separate *I* and *Q*

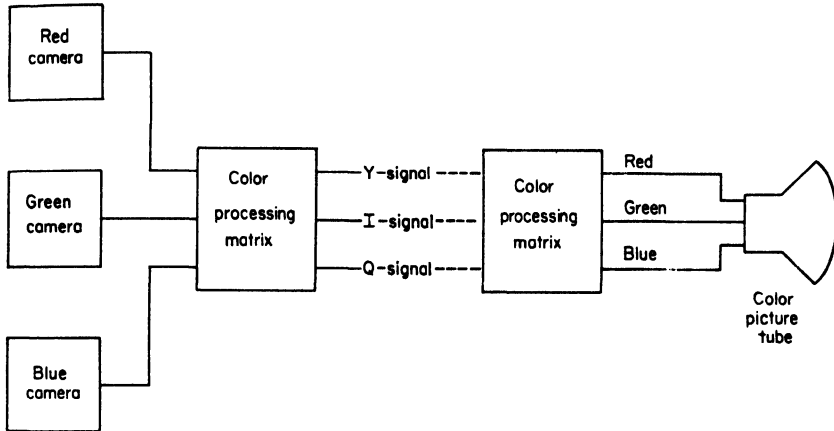



FIG. 67. Method by which camera-output signals are processed for transmission and restored in reception.

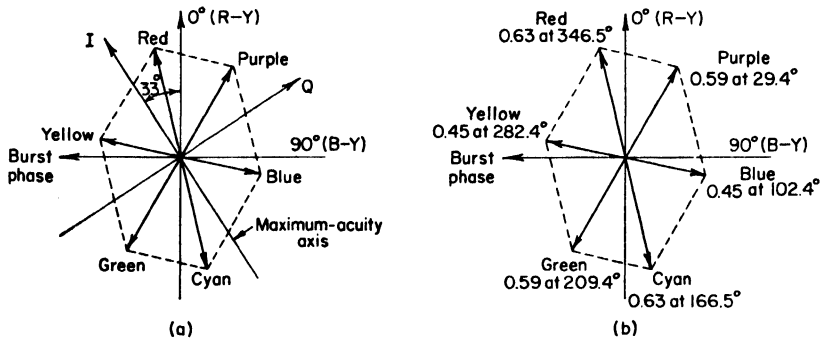



FIG. 68. (*a*) Vector diagram showing axis corresponding to maximum acuity for human vision. (*b*) Subcarrier phases and amplitudes for six colors.

signals. This is illustrated in the simplified diagram in Fig. 67. At the receiver a corresponding matrix may be employed for restoring the red, green, and blue signals.

The *I* and *Q* vectors, when modulated, may be considered to produce a resultant vector sum. These vectors may be reversed in polarity. Thus, in combination, the sum of these vectors may have any phase from 0 to 360 deg. The phase corresponds to the color, or hue, of the signal transmitted and received. The amplitudes of the *I* and *Q* signals which vary separately produce changes in the amplitude of the vector sum. The amplitude is a function of the saturation of the color transmitted and received.

Figure 68 shows the group of colors produced by the FCC approved signal-standards color system when the phase of the reference color burst signal properly corresponds with

the correct axis as shown. The I signal is fixed on the orange-cyan axis and the Q signal, lagging 90 deg, is fixed on the purple-yellowish-green axis. The orange-cyan axis was selected for the I signal to take advantage of the spectral sensitivity of the human eye which is most acute for these colors. Fine detail may be most clearly detected on this axis. The band width of the I-signal channel was established to conform with the amount of detail the eye can distinguish on this axis under normal viewing conditions. It is specified as follows:

At 1.3 Mc less than 2 db down
At 3.6 Mc at least 20 db down

The useful band width is approximately 1.5 Mc.

The eye has less acuity for colors on the Q-signal axis and the specification is as follows:

At 400 kc less than 2 db down
At 500 kc less than 6 db down
At 600 kc at least 6 db down

The useful band width is approximately 500 kc.

These specifications accomplish two important objectives. First, they provide adequate color detail for those colors to which the eye is most sensitive and restrict the band width to useful dimensions for colors to which the eye is insensitive in fine detail. Thus, the minimum necessary band width is employed. Second, they make it possible to utilize vestigial-side-band transmission of the 1.5-Mc I signal in combination with a relatively high subcarrier frequency, without cross talk with the Q channel.

Attention is directed to the fact that the subcarrier frequency of 3.579545 Mc is only 0.57 Mc below the frequency of 4.15 Mc at which the video response of the receiver must cut off. Two narrow color channels of about 0.5 Mc could be employed with symmetrical side-band transmission without cross talk; but, if the channels were increased in band width to the point where the upper side bands were attenuated, cancellation of the quadrature side bands between the I and Q channels would not take place and cross talk would result. By restricting the band width of the Q channel to 0.5 Mc so that side-band attenuation does not occur and the unsymmetrical side bands from the I signal are rejected, the wide-band I channel may be employed without cross talk.

The I signal may be obtained in a roundabout manner, not ordinarily employed, from the vector sum of signal vectors from the red and blue cameras. After the brightness component is removed to produce the $E_R - E_Y$ and the $E_B - E_Y$ color-difference signals, they are combined to produce the E_I signal in the following proportions.

$$E_I = -0.27(E_B - E_Y) + 0.74(E_R - E_Y) \qquad (35)$$

It will be noted that the polarity of the $E_B - E_Y$ signal was inverted to produce the proper I-signal phase of 57 deg corresponding to orange. Inversion of the phase is accomplished by passing the signal through a plate-coupled amplifier stage. Apportioning of the red and blue color-difference signals is accomplished by voltage-dividing resistors. The inverted $E_B - E_Y$ signal corresponds to the color which is complementary to blue, which is approximately yellow.

Following the creation of the I signal, it is passed through a filter which limits the band width to about 1.5 Mc, as explained previously.

The Q signal by this method would be correspondingly produced by combining certain proportions of the red and blue color-difference signals. Since the Q vector lies between the red and blue vectors, no inversion is necessary. The Q-signal specification is as follows:

$$E_Q = 0.41(E_B - E_Y) + 0.48(E_R - E_Y) \qquad (36)$$

The vector sum of these vectors corresponds to the color reddish-blue or purple. The color complementary to this, yellowish-green, is obtained by reversing the polarity of the Q signal which is accomplished by reversing the polarities of the red and blue color-difference signals of which it is composed. The Q signal is passed through a filter which limits its band width to about 0.5 Mc.

Figure 69 shows how the video spectrum is utilized in the simultaneous transmission of the luminance signal, the vestigial-side-band I signal, and the double-side-band Q signal.

The equations for the I and Q signals were shown, in terms of the red and blue color-difference signals, to be

$$I = 0.74(R - Y) - 0.27(B - Y) \qquad (37)$$

$$Q = 0.48(R - Y) + 0.41(B - Y) \qquad (38)$$

The equation for the brightness signal was shown to be

$$Y = 0.30R + 0.59G + 0.11B \tag{39}$$

It may be proved that the green component is present by substituting in the I and Q equations the color values representing the brightness signal. Solving these equations gives the following components for the I and Q signals, which are equivalent to Eqs. (37) and (38):

$$I = 0.60R - 0.28G - 0.32B \tag{40}$$
$$Q = 0.21R - 0.52G + 0.31B \tag{41}$$

This operation illustrates that the I and Q signals may be produced at the camera output without the intermediate step of producing the red and blue color-difference signals. This is the simpler method conventionally employed. It may also be shown that the red and blue color-difference signals may be obtained by cross-mixing the I and Q demodulator outputs, by solving the I and Q equations for the red and blue color-difference signals expressed as components of I and Q. They are

$$B - Y = 1.7Q - 1.1I \tag{42}$$
$$R - Y = 0.62Q + 0.96I \tag{43}$$

In practice, all three color-difference signals may be abstracted from the received signal by appropriate demodulation.

The complete color-picture signal has the following composition:

$$E_M = E_Y' + [E_Q' \sin(\omega t + 33 \text{ deg}) + E_I' \cos(\omega t + 33 \text{ deg})] \tag{44}$$

where E_Y', E_Q', and E_I' are the gamma-corrected signals, E_M is the total video voltage corresponding to the scanning of a particular picture element, and ω is 2π times the subcarrier frequency.

Fig. 69. Simultaneous utilization of video spectrum by luminance, I, and Q channels in FCC Standards for Compatible Color Television.

For color-difference frequencies below 0.5 Mc, the signal may be represented by

$$E_M = E_Y' + \left\{\frac{1}{1.14}\left[\frac{1}{1.78}(E_B' - E_Y')\sin\omega t + (E_R' - E_Y')\cos\omega t\right]\right\} \tag{45}$$

82. Color-burst Phase. It will be noted from Fig. 68 that the phase of the color burst differs from that of the I-signal axis by 57 deg and not 90 deg, and from the Q-signal axis by 147 deg instead of 180 deg. Tests showed that certain burst phases produced minimum visibility in the return trace but simultaneously they also adversely affected d-c restoration. No compelling reasons exist for rotating the burst phase from 57 deg, but an advantage would accrue in cost savings in manufacturing inexpensive color receivers which did not take full advantage of the maximum possible color detail provided for in the standards. In such inexpensive receivers, for example, the $R - Y$ and $B - Y$ signals (instead of I and Q) could be used, each having a band width sufficiently narrow (not more than about 0.5 Mc) so that matrixes would be simplified and cheapened, and cross-talk problems resulting from vestigial I-signal side-band transmission could be eliminated.

83. Delay Circuits. From the output of the color cameras the Y signal is produced directly by combining the red, green, and blue voltages in proper proportion in a simple resistance network. However, the I and Q signals pass through band-limiting filters. When a signal passes through a limiting filter, it is delayed in time. The delay is greatest when the band limiting is greatest. Therefore, unless corrective steps were taken the Y signal at the receiver would arrive first, followed next by the I signal and last by the Q signal. When combined and viewed on a picture tube, the color information would be displaced from the brightness information in the horizontal plane and poor vertical edges would result. To cause the Y, I, and Q signals to coincide in arrival time, delay networks are added in the I and Y channels at the transmitter to equalize the delays. The delay introduced in the Y channel is about 1 μsec.

84. Gamma Correction. The linear relationship of the light-input values at the camera tube and the light-output values at the receiver tube is one measure of the

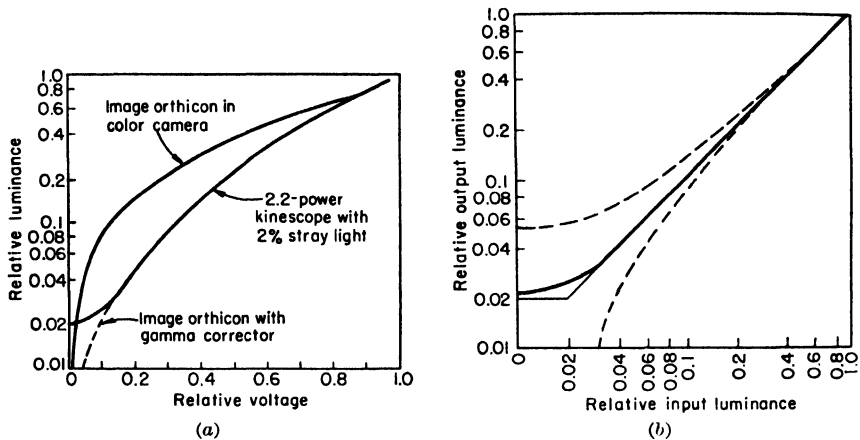



FIG. 70. (a) Plot comparing a color-camera image orthicon curve with that of a 2.2-power kinescope with 2 per cent stray light. (b) Over-all transfer characteristics for a corrected-color-image orthicon camera driving a 2.2-power color kinescope with 2 per cent stray light. The slight "toe" on the characteristic results from failure to achieve an infinite slope in the gamma corrector. Dotted lines show curves for 5 per cent and zero stray light conditions with no change in relative bias.

fidelity of the system but in monochrome transmission considerable tolerance has been permissible. In color transmission the situation is quite different, because reproduced color fidelity depends upon the combinations of three voltages from three separate color-camera tubes which operate at random over wide amplitude ranges. There must be gamma correction to compensate for the nonlinear relationship between input voltage and light output of receiver picture tubes, as in monochrome television. But there must also be a close match between the transfer characteristics of the three color channels. A difference in the response between these channels destroys the relationships of the voltages which must combine in the correct proportions for true color reproduction. In addition, this may produce incorrect color values at various places on the brightness or saturation range.

By a wide margin the most essential phase of gamma correction is obtaining equality in the three primary color channels at all operating amplitudes. Lack of it produces noticeable and objectionable errors in chromaticity and will be especially objectionable when white should be reproduced.

Gamma correction is an essential part of the color-transmitting system and is accomplished in gamma-correcting circuits at the outputs of the red, green, and blue

camera tubes. Figure 70*a* from Wentworth compares the characteristics of a color-camera image orthicon tube and a 2.2-power kinescope with 2 per cent stray light. Figure 70*b* shows the over-all gamma-corrected characteristic.

85. Phase Distortion and Compensation. The FCC standards specify a transmission frequency response characteristic and a corresponding receiver characteristic as shown in Fig. 41. In this system there are sources of phase distortion and associated envelope delay. Unless compensation is introduced, the envelope delay through the system is sufficient to produce noticeable picture degradation.

The sharp receiver cutoff at about 4.1 Mc produces phase shift which is not proportional to frequency and becomes objectionable above about 2.5 Mc. The resulting overshoots following sharp vertical edges produce ringing and smear. This effect is minimized by phase correction which distributes the overshoot in a sym-

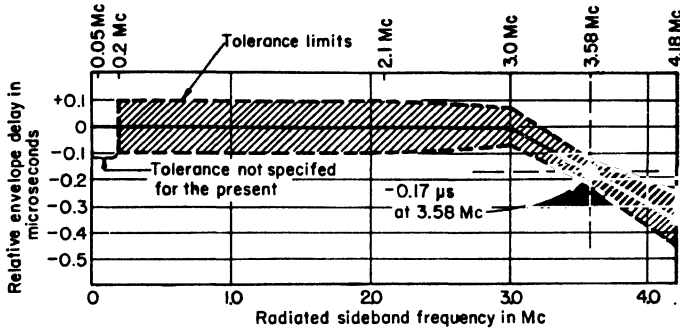



FIG. 71. Envelope delay specified to compensate for vestigial-side-band filter and receiver high-frequency cutoff.

metrical manner about the transition point and reduces the amplitude about 50 per cent.

The vestigial-side-band filter attenuation produces envelope delay which is decreasingly significant at frequencies up to about 3 Mc. This shows as a leading undershoot and loss of sharpness of vertical edges. Compensation minimizes the undershoot and resulting smear.

The United States standards specify that a sine wave introduced at those terminals of the transmitter which are normally fed the composite color-picture signal shall produce a radiated signal having an envelope delay, relative to the average envelope delay between 0.05 and 0.2 Mc, of zero microseconds up to a frequency of 3 Mc and then linearly decreasing to 4.18 Mc, so as to be equal to -0.17 μsec at 3.58 Mc. The tolerance on the envelope delay shall become ± 0.1 μsec at 2.1 Mc and remain at ± 0.1 μsec down to 0.2 Mc. The tolerance shall also increase linearly to ± 0.1 μsec at 4.18 Mc. This specification is illustrated in Fig. 71. The tolerance of ± 0.05 is equal to about one-half of a horizontal picture element.

The over-all color-transmitting system incorporating the elements which have been described is illustrated in simplified manner in Fig. 72.

86. Color-television Receivers. Many variations of circuitry are possible in receivers designed to United States color standards. Figure 73 shows, for example, various circuit arrangements that have been used. The first detector is omitted from the diagram for simplicity. From the second detector one circuit drives the sound system similar to those in monochrome receivers. Another circuit from the second detector may drive a video amplifier No. 1, having two outputs. One contains a delay line for correction of time delay of the luminance signal. This output does not transmit the 3.579-Mc subcarrier. It contains a contrast control and supplies brightness information to the adders. The other output from video amplifier No. 1

contains another contrast control mechanically actuated on a common shaft with the one referred to. This control drives a band-pass chroma amplifier, No. 2, which may be band-limited to 1.5 Mc in the chroma output. A chroma-voltage control from this output drives the *I* and *Q* demodulators.

From amplifier No. 1, 3.579-Mc subcarrier burst voltage is fed to a burst amplifier, phase detector, reactance tube, and subcarrier reference oscillator which in combination lock the phase of the latter unit to that of the subcarrier oscillator at the transmitting point of origination. The color subcarrier reference oscillator utilizes only the burst portion of the chroma signal. Burst gate pulses from the horizontal deflection system admit the bursts but block out all other components of the incoming signal

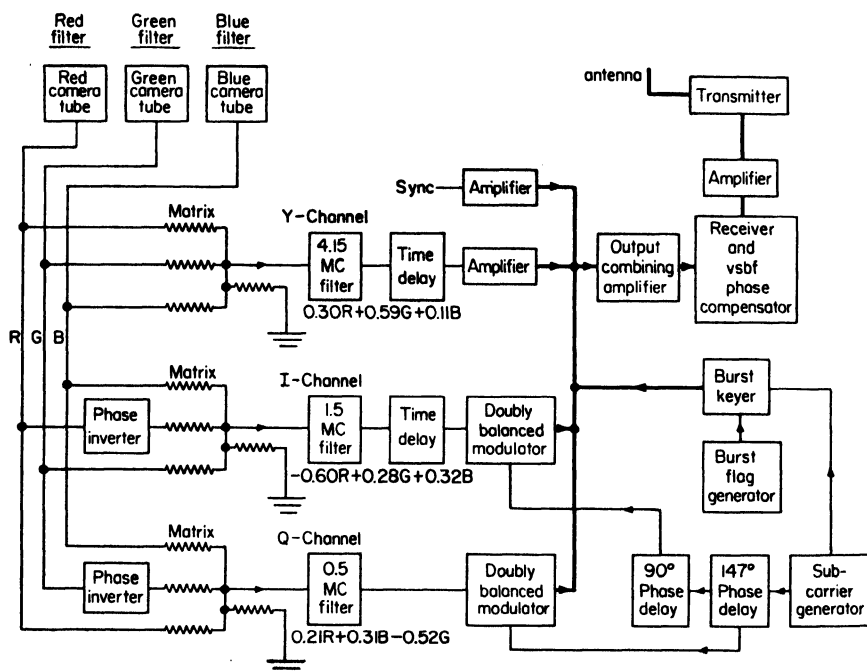



FIG. 72. Elements of a color-transmitting system.

at the burst amplifier. From the reference oscillator subcarrier voltage is fed directly to the *I* synchronous detector. Also, through a 90-deg phase shifter, subcarrier voltage is fed to the *Q* synchronous detector. The *Q* channel, as shown, is band-limited by a 0.5-Mc low-pass filter and the *I* channel is similarly limited to 1.5 Mc.

During transmission of monochrome pictures it is desirable to disable the color channels so that only the output of the brightness channel is present in the adders. This disabling then eliminates noise or high-frequency monochrome signal components from the adders where they could generate spurious low-frequency color-voltage components. The disabling is accomplished by a color killer. When burst voltage is received it provides a voltage which overcomes the cutoff bias voltage otherwise present on amplifier No. 2.

The brightness and synchronizing deflection circuits are very similar to those in monochrome receivers. One circuit may provide AGC voltage for the r-f and i-f systems by deriving the AGC voltage from the horizontal deflection system, as shown. This method of AGC provides the close distribution of control which is more important in color receivers than monochrome.

A color receiver essentially differs from a monochrome receiver by the addition of color-processing components which have no counterpart in monochrome receivers. These components, as represented in Fig. 73, are the *I* and *Q* quadrature color channels, the reference subcarrier oscillator and its auxiliaries, the matrixes, and the adders which combine the brightness and the color voltages.

Two important alternatives in color-receiver detector design are employed. In one, as illustrated in Fig. 73, the brightness, color, and sound signals are taken from a common second detector by means of v-f selective circuits. In the other, the last i-f stage or stages contain branches with two or three second detectors, each of which is followed by its own selective circuits and amplifiers. A single detector may function for the brightness and sound signals or, in alternative designs, for chrominance and sound.

FIG. 73. Principal components of circuitry of one form of a color-television receiver following the second detector. Main signal channels have heavy lines for easy identification.

Receivers have been used commercially which take full advantage of the 1.5-Mc *I*-signal band width. However, as indicated previously, the incoming chrominance information may be band-limited in each channel to 0.5 Mc or less into the $R - Y$ and $B - Y$ signals, at some reduction in cost, circuitry, and color detail. In either case the outputs of the matrixes are $R - Y$, $G - Y$, and $B - Y$.

It is desirable to maintain the superheterodyne local oscillator frequency within 50 kc when employing a single brightness-color detector to keep the sound-channel frequency in the minimum response area of the sound-suppression notch. Excessive frequency drift will increase the amplitude of the sound voltage in the picture circuits and correspondingly increase the amplitude of the sound-subcarrier beat note of 920 kc. This cannot be eliminated in the brightness amplifiers. It must be controlled by reducing the sound i-f level. It is proportional to the product of the i-f chrominance subcarrier and the i-f sound carrier. For satisfactory suppression of the beat voltage it is customary to maintain it 40 db below the brightness-signal level.

References

Abrahams, I. C.: "Choice of Chrominance Subcarrier in the NTSC Standards," NTSC Technical Monograph 9.

———: The Frequency Interleaving Principle in the NTSC Standards, *Proc. IRE*, January, 1954.

Applebaum, S.: Gamma Correction in Constant Luminance Color TV Systems, *Proc. IRE*, **40**, October, 1952.

Bailey, W. F.: The Constant Luminance Principle in Color TV, *Proc. IRE*, January, 1954.

——— and C. J. Hirsch: "Quadrature Crosstalk in NTSC Color TV," NTSC Technical Monograph 2-A.

Bedford, A. V.: Mixed Highs in Color TV, *Proc. IRE*, **38** (9), September, 1950.
Bingley, F. J.: "Colorimetry in Color TV," Parts I, II, and III, NTSC Documents P12-220, P12-297, P12-344.
Brown, G. H.: NTSC Technical Monograph 14, p. 12.
———: Choice of Axes and Bandwidths for the Chrominance Signals in NTSC Color TV, *Proc. IRE*, January, 1954.
———: NTSC Technical Monograph No. 4, NTSC Document p. 12.
——— and D. G. C. Luck: Principles and Development of Color TV Systems, *RCA Rev.*, **14** (2), June, 1953.
Engstrom, E. W.: Basic Concepts and Evolution of Color TV, *Proc. IRE*, January, 1954.
Epstein, D. W.: Colorimetric Analysis of RCA Color TV Systems, *RCA Rev.*, **14** (2), June, 1953.
Feingold, W. F.: Matrix Networks for Color TV, *Proc. IRE*, January, 1954.
Fredendall, G. L.: Delay Equalization in Color TV, *Proc. IRE*, January, 1954.
Janes, R. B., L. B. Headrick, and J. Evans: Recent Improvements in the 21 AX P22 Color Kinescope, *RCA Rev.*, June, 1956.
Kell, R. D., and A. C. Schroeder: Optimum Utilization of the Radio Frequency Channel for Color Television, *RCA Rev.*, **14** (2), June, 1953.
Livingston, D. C.: Colorimetric Properties of Gamma Corrected Color TV Systems, 1953 *IRE Convention Record*, Part 4.
MacAdam, D. L.: Quality of Color Reproduction, *Proc. IRE*, **39**, May, 1951.
Palmer, R. C.: System Delay Characteristic in NTSC Color TV, *Proc. IRE*, January, 1954.
Pritchard, D. H., and R. N. Rhodes: Color TV Signal Receiver Demodulators, *RCA Rev.*, **14** (2), June, 1953.
Richman, D.: Theory of Synchronization Applied to NTSC Color Television, 1953 *IRE Convention Record*, Part 4, March, 1953.
Rotow, A. A., R. S. Veith, and R. G. Neuhauser: Image Orthicons for Color TV Cameras, *Proc. IRE*, January, 1954.
Wentworth, John W.: "Color Television Engineering," McGraw-Hill.
———: *Broadcast News*, January, 1954.
Wintringham, W. T.: Color TV and Colorimetry, *Proc. IRE*, **39** (10), October, 1951.
NTSC Signal Specifications, *Proc. IRE*, January, 1954.
NTSC Color TV Standards, *Proc. IRE*, January, 1954.

87. Tricolor Kinescope. The RCA type of shadow-mask color-picture tube is used in all color receivers now produced. Three electron guns are employed, one each for red, green, and blue. The electron streams converge at a perforated masking plate, as shown in Fig. 74*a* and *b*. After passing through a perforation the beams diverge and impinge upon separate phosphor dots which, when so excited, produce the red, green, and blue light corresponding to the individual exciting electron beams. One set of red-, green-, and blue-emitting phosphor dots is located in a triangular configuration behind each perforation. Current tubes contain 356,666 perforations and sets of dots, representing a total of 1,070,000 dots. The three electron beams are individually focused and, by an electrostatic-lens system, are made to converge at the aperture mask or screen. They are electromagnetically deflected in the horizontal and vertical planes simultaneously by a common yoke. Convergence at a common point is produced by separate convergence controls.

Dynamic convergence controls are also provided. The reason is illustrated in an exaggerated degree by reference to Fig. 74*c*, which shows an earlier type of color kinescope having a flat shadow-mask plate. An adjustment which produced correct convergence control at point *A* would not have been correct for point *B*. Therefore, dynamic convergence control is provided to correct for the changing focal distance as the beam travels from top to bottom and across the plate. This is accomplished by the application of parabolic wave forms to the convergence circuits, obtained from the horizontal- and vertical-deflection circuits. Although the modern curved masked kinescope shown in Fig. 74*a* has more nearly uniform focal distance as the beam traverses, it is not entirely uniform. Figure 74*c* is included only to illustrate in exaggerated form the effect described.

The shadow-mask holes and dots are so positioned that the electron beam from, for example, the green gun can strike only green-emitting dots and, similarly, the red and blue beams strike only red- and blue-emitting dots. The intensity of the light from the individual dots is controlled by the beam intensities, which are in turn controlled by the individual gun grids. The eye automatically integrates the colors and their intensities to recreate the original picture content. White light is observed by the eye at any reasonable viewing distance when the proper ratio of red-, green-, and

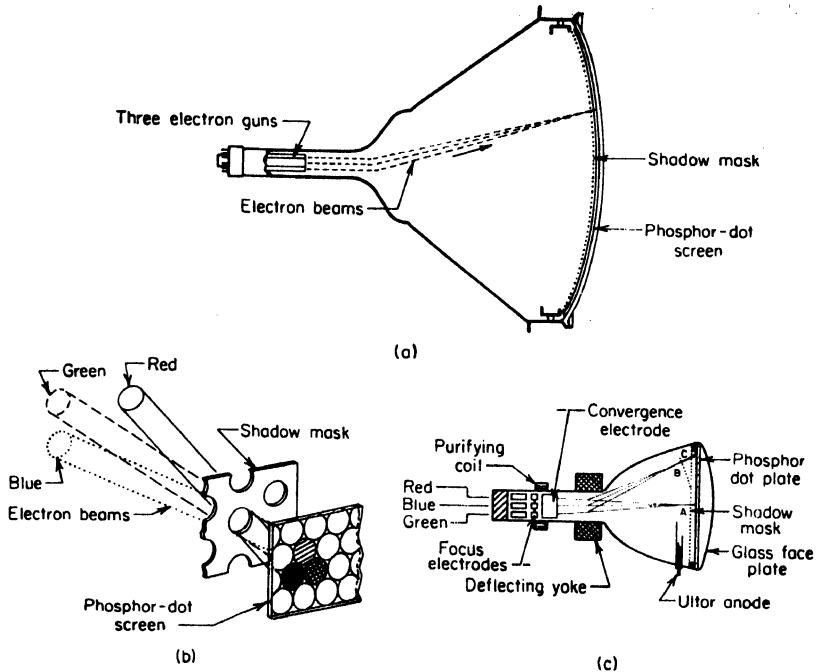



FIG. 74. Shadow-mask color kinescope.

blue-dot excitation is produced, despite the fact that close inspection shows only red, green, and blue dots.

TEST AND ALIGNMENT PROCEDURES

Faithful transmission and reproduction of television pictures require that the adjustment of the components of the complex system be correct. Standard test procedures are necessary so that the performance of the system or its components may be evaluated simply and rapidly. Certain fundamental parameters require special attention.

88. Amplitude Response vs. Frequency. The gain of the system should be uniform over the v-f range. In color transmission in particular a difference in gain in the area of the color subcarrier will change the amplitude of the color components compared with the brightness signal and the reproduced color will be too low or too high in saturation. To evaluate this characteristic of the system a multifrequency burst signal may be used. Such a generator produces its own line-frequency blanking and synchronizing signal. No vertical signal is necessary. The trailing edge of the blanking signal triggers a chain of six multivibrators which function in sequence to introduce or key-in during each line interval a choice of six different video frequencies. These are transmitted in succession at equal amplitude and are reproduced at the terminal point on an oscilloscope. The output may be observed visually and may be photographed on a Land camera from which the complete print is available in 1 min. The wave form of the burst signal is shown in Fig. 75. Built-in oscillators are used to generate each of the six frequency bursts and each of these oscillators is variable over a range of frequencies. The sequence may be, for example, 0.5, 1.5, 2.0, 3.0, 3.6, and 4.2 Mc. An oscilloscope is used not only at the terminal point but also at the transmitting point to monitor the signals and adjust levels. The response of the system at each frequency may be compared with the blanking and sync signal

and with the other frequencies, and also with a white bar which is transmitted just prior to the first sine-wave burst. Nonlinear amplitude response may be observed when present at any burst frequency. Harmonic distortion may be observed by expanding the oscilloscope scale, particularly at the lowest burst frequencies.

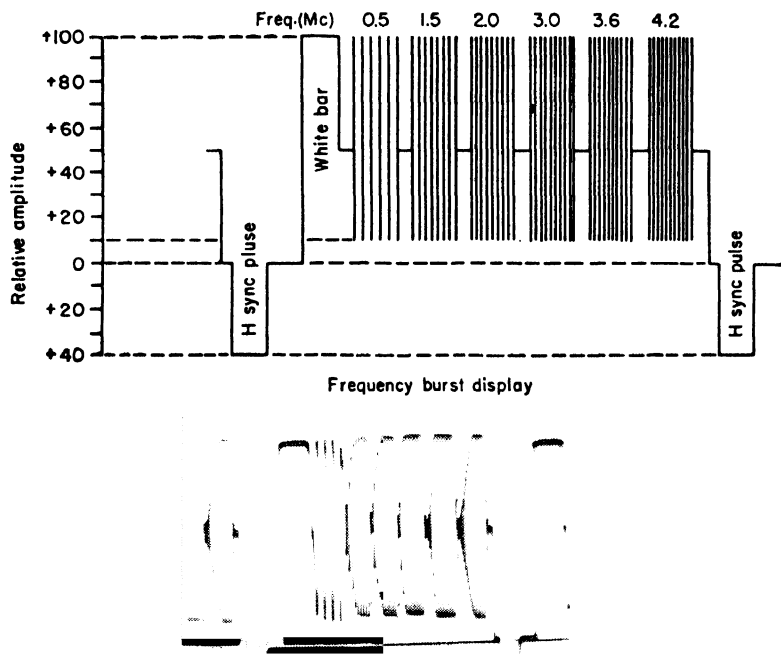



FIG. 75. Multiburst signal and CRO wave-form display.

89. Smear and Streaking. Streaking is produced by irregular frequency responses, usually at the low and middle video frequencies. Because of the system clamping circuits it is seldom present below 60 cps and in practice it is usually below 200 kc, at line frequency or multiples of it. This irregularity may be observed on an oscilloscope by transmitting and observing a white window signal such as is shown in Fig. 76. It may also be observed on a picture monitor. Streaking may be positive when a white streak follows a white image, or negative when a white streak follows a black image, or it may be a combination of both. With the oscilloscope adjusted for the horizontal rate, streaking will appear as low-frequency undershoots or overshoots to the rectangular white window signal and these usually may be seen to be sine waves or parts of sine waves added to the horizontal flat top of the window or the setup line. By observing the horizontal interval of the sine-wave addition, its frequency may be estimated by comparing it with the line frequency of 15,750 kc. The window signal is produced by a generator containing adjustment for control of window size. To produce sensitivity to low-frequency irregularities, the window is often made large, about half of picture height and width. A video resonant circuit in the transmitting system will produce ringing which will be visible on picture tubes as a series of closely spaced images displaced to the right of vertical lines. The amplitude and frequency of the irregularity may be evaluated closely by observing the window signal on an oscilloscope. The window signal contains both horizontal and vertical sync signals.

90. Amplitude and Phase Linearity. A stair-step signal is used to test and evaluate amplitude and phase linearity, which are referred to as differential gain and differential phase response. Usually 10 equal steps are transmitted ranging progressively from black level to white level as shown in Fig. 77. In reproduction, differences in the amplitudes of the steps indicate nonlinear amplitude response. Superimposed upon each step there may be sine waves of color subcarrier frequency having on each step an amplitude of about twice the step amplitude. If the latter composite signal is passed through a filter which eliminates the frequencies below the subcarrier, the subcarrier components of each step are then lined up horizontally on the same axis. A difference in amplitude between the 10 subcarrier segments, now shown side by

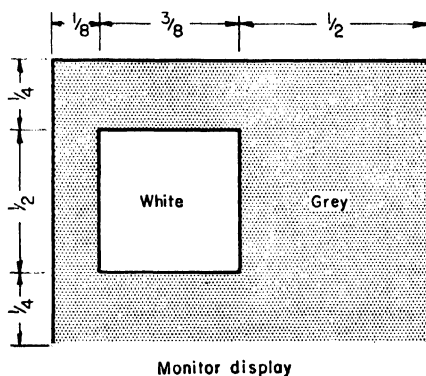



Fig. 76. White-window signal and CRO display.

side, indicates that the gain of the system differs over the amplitude range, and shows also the portion of the range where it exists.

The measurement of differential phase is made by comparing the phase of the individual subcarrier segments with a reference phase. A calibrated phase shifter is employed to measure the degree of phase departure, by observing the amount of phase shift required to make the reference phase coincide with that of the subcarrier segment being measured.

91. NBC Dynamic Test Signal. The test signals previously described cannot be used during program transmission. There has been a growing need for a standard signal which could be transmitted during program periods, without interference, to measure and establish levels and measure the system response. Such a signal has been under development and is in experimental operation. It may undergo substantial revision before being accepted as an industry standard and in its present form represents only one possible form of signal.

This standard signal may occupy from three to six lines near the end of the vertical blanking. Three lines are normally transmitted but the number and the exact

location are tentative and subject to additional tests and observations. The lines of the standard test signal are identical in form. The vertical-blanking signal and color-reference-burst signal are assumed to be transmitted normally. The test signal being described is added in each line following the reference burst signal.

With respect to time, following the back porch, there is a bar having an amplitude of 50 units on the IRE scale and a duration of 20 μsec. The amplitude of this bar is precisely adjusted and maintained. Superimposed an dcentered on the bar is 10 μsec of a 3.579-Mc sine wave phased tentatively 180 deg from the reference sub-carrier burst. At the conclusion of the bar a 2-μsec interval occurs at blanking level

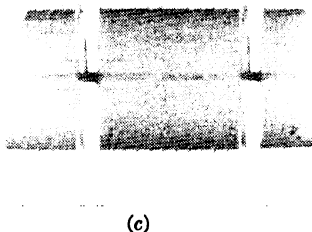
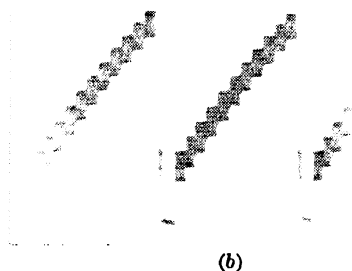
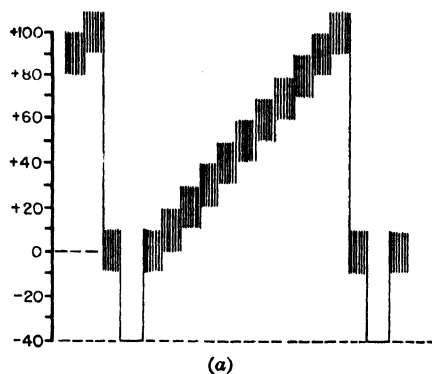



FIG. 77. Stair-step signal. (a) CRO wave-form display. (b) Normal signal. (c) Signal after h-p filter.

whereupon another 3.579-Mc sine-wave burst is transmitted, phased for magenta, 60.8 deg, of a duration of 10 μsec. This is followed by 2 μsec of blanking level, whereupon another bar appears at 100 on the IRE scale of a duration of 20 μsec. Centered in this bar in time there is a 10-μsec interval of 3.579-Mc sine wave phased for cyan, 283 deg, the amplitude of which varies between 60 and 100 on the IRE scale. Following this bar the axis returns to the blanking level and the horizontal synchronizing pulse for the next frame appears in conventional manner. The amplitudes of the sine waves span 40 deg on the IRE scale in each case. The experimental standard test signal and the significant phases are illustrated in Fig. 78.

Monochrome Standard Signal. For monochrome transmissions the 3.579-Mc sine waves are removed but the pedestal levels remain. By observing the level of these pedestals at the output of a system, it is possible to determine if amplitude compression exists, and the degree.

Color Standard Signal: For color transmissions the above comments apply in addition to the following. The amplitudes of the three sine-wave signals are the same as the color burst in the video signal. These are also the same as sync amplitude.

Thus sync, burst, and chroma as well as white levels are all established by the standard signal.

By passing the signal through the conventional high-pass filter for this purpose all the low-frequency components are removed and as a result the a-c axis of color burst and the three sine waves of the standard signal are in line. If no differential gain distortion is present in the system these should all have the same amplitudes.

When displayed on a vectorscope a color program combined with the standard signal will contain three dots indicating subcarrier burst minus 180 deg, cyan, and magenta which may be clearly seen. Thus differential phase shift is easily measured

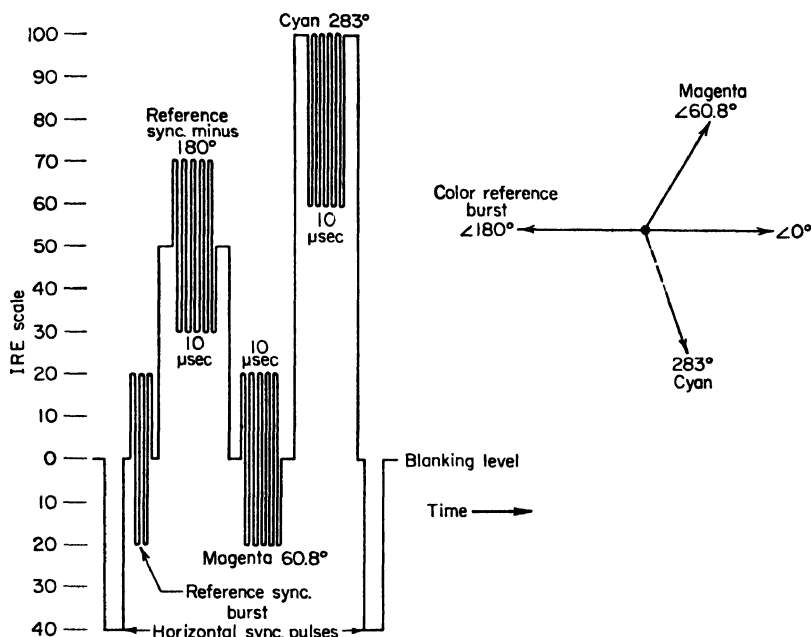



FIG. 78. Single line of NBC experimental standard test signal transmitted during vertical blanking interval, following vertical sync. Three consecutive lines are transmitted.

while the system is in use. The three lines of information provided by the standard signal per field are adequate for a vectorscope presentation.

It is expected that the standard signal will be inserted so that it appears as early in the system as possible. This means the system from camera to receiver may be measured.

It should be pointed out that either adjusting vertical centering in a monitor or receiver or a slight misadjustment of the vertical-hold control will permit the three bars of color on their respective pedestals to appear above the top of the picture. These signals may aid in adjusting a color receiver.

This signal promises to make local and network operations much simpler and more consistent and enable the telephone company to locate circuit faults more easily. Performance characteristics of transmitters may be more accurately determined by its aid.

References

Color Television:

Barstow, J. N.: The ABC's of Color TV, *Proc. IRE*, November, 1955.

Cherry, W. H.: Colorimetry in Television, *RCA Rev.*, **8**, 427, September, 1947.

Clapp, R. G.: Color Receiver Design, A Review of Current Practice, *Proc. IRE*, **44** (3), March, 1956.

Fink, D. G.: Color Television on Ultra-high Frequencies, *Electronics*, **19**, 109, April, 1946.
———: Two Systems of Color Television, *Electronics*, **20**, 72, January, 1947.
———: Where Color Television Stands, *Electronics*, **19**, 104, May, 1946.
Fisher, Joseph F.: Generation of NTSC Color Signals, *Proc. IRE*, **41** (3), March, 1953.
Gibson, W. G., and A. C. Schroeder: Color TV Luminance Detail Rendition, *Proc. IRE*, August, 1955.
Gluyas, T. M.: TV Transmitter Considerations in Color Broadcasting, *RCA Rev.*, September, 1954.
Goldmark, P. C., *et al.*: Color Television, *Proc. IRE*, **30**, 162, April, 1942; **31**, 465, September, 1943.
Kell, Fredendall, Schroeder, and Webb: An Experimental Color Television System, *RCA Rev.*, **7**, 141, June, 1946.
——— *et al.*: An Experimental Simultaneous Color-television System (in three parts), *Proc. IRE*, 861, August, 1947.
RCA Laboratories: Simultaneous All-electronic Color Television, *RCA Rev.*, **7**, 459, December, 1946.
Color Television Issue, *Proc. IRE*, October, 1951.
Color Television Issue, *Proc. IRE*, January, 1954.

General Bibliography

Books:

Fink, D. G.: "Principles of Television Engineering," McGraw-Hill, 1940.
———: "Television Standards and Practice," McGraw-Hill, 1943.
——— (ed.): "Television Engineering Handbook," McGraw-Hill Book Company, Inc., New York, 1957.
Lohr, L. R.: "Television Broadcasting," McGraw-Hill, 1940.
Maloff, I. G., and D. W. Epstein: "Electron Optics in Television," McGraw-Hill, 1938.
Wilson, J. C.: "Television Engineering," Pitman, 1937.
Zworykin and Morton: "Television," Wiley, 1940.

Periodicals:

Beal, R. R.: Equipment Used in the Current RCA Television Field Tests, *RCA Rev.*, **1** (3), 36, January, 1937.
Beers, Engstrom, and Maloff: Some Television Problems from the Motion Picture Viewpoint, *J. Soc. Motion Picture Engrs.*, **32**, 18, January, 1939.
Bingley, F. J.: VHF Multiple-relay Television Network, *Electronics*, **18**, 102, October, 1945.
Bloecker, W. E.: Interconnecting Facilities for Television Broadcasting, *Electronics*, **20**, 102, November, 1947.
Boothroyd, W.: U-h-f Television Relay System, *Electronics*, **20**, 86, March, 1947.
Bradley and Traub: A New Television Projection System, *Electronics*, **20**, 84, September, 1947.
Buzalski, T. J.: A Method of Measuring the Degree of Modulation of a Television Signal, *RCA Rev.*, **7**, 265, June, 1946.
Cowan, F. A.: Coaxial Cable Networks, *Proc. IRE*, **35**, 1364, November, 1947.
Dearing, Gihring, Guy, and Kear: Multiple Television and FM Transmitting Antenna Installation on the Empire State Building, *Proc. IRE*, **41** (3), March, 1953.
Deerhake, F. M.: 2000-Mc Television Program Chain, *Electronics*, **21**, 94, February, 1948.
DuMont and Goldsmith: Television Broadcast Coverage, *Proc. IRE*, **32**, 192, April, 1944.
Eddy, W. C.: Television Lighting, *J. Soc. Motion Picture Engrs.*, **33**, 41, July, 1939.
———: Television Studio Considerations, *Communication and Broadcast Eng.*, **4** (4), 12, April; **5**, 14, May; **6**, 20, June; **7**, 17, July, 1937.
Engstrom, Kell, Bedford, Trainer, Holmes, Carlson, and Tolson: An Experimental Television System, *Proc. IRE*, **22**, 1241, 1246, 1266, November, 1934.
Epstein and Pensak: Improve Cathode-ray Tubes with Metal-backed Luminescent Screens, *RCA Rev.*, **7**, 5, March, 1946.
Fink, D. G.: Brightness Distortion in Television, *Proc. IRE*, **29**, 310, June, 1941.
———: Design Trends in Television Transmitters, *Electronics*, **21**, 76, January, 1948.
———: A Laboratory Television Receiver (in six parts), *Electronics*, Part I, *ibid.*, **11** (7), 16, July, 1938; Part II, *ibid.*, **11** (8), 26, August, 1938; Part III, *ibid.*, **11** (9), 22, September, 1938; Part IV, *ibid.*, **11** (10), 16, October, 1938; Part V, *ibid.*, **11** (11), 26, November, 1938; Part VI, *ibid.*, **11** (12), 16, December, 1938.
———: Photographic Analysis of Television Images, *Electronics*, **14**, 24, August, 1941.
———: Television Antennas for Apartments, *Electronics*, **20**, 96, May, 1947.
———: Television Broadcasting Practice in America, 1927 to 1944, *J. IEE (London)*, **92**, III, 145, September, 1945.
———: A Television Receiver for the Home, *Electronics*, **12** (9), 16, September, 1939.
Friend, A. W.: Television Deflection Circuits, *RCA Rev.*, **8**, 86, March, 1947.
Gannett and Green: Wire Transmission System for Television, *Bell System Tech. J.*, **6**, 616, October, 1927.
Goldmark, P. C.: A Continuous Type Television Film Scanner, *J. Soc. Motion Picture Engrs.*, **33**, 18, July, 1939.
———: Problems of Television Transmission, *J. Applied Phys.*, **10**, 447, July, 1939.
——— and Dyer: Quality in Television Pictures, *Proc. IRE*, **28**, 343, August, 1940.
Gray, Horton, and Mathes: The Production and Utilization of Television Signals, *Bell. System Tech. J.*, **6**, 560, October, 1927.
Hanson, O. B.: Experimental Studio Facilities for Television, *RCA Rev.*, **1** (4), 3, April, 1937.
Ives, H. E.: Television, *Bell System Tech. J.*, **6**, 551, October, 1927.

———: Transmission of Motion Pictures over a Coaxial Cable, *J. Soc. Motion Picture Engrs.*, **31**, 256, September, 1938.

Kallman, H. E.: Gradation of Television Pictures, *Proc. IRE*, **28**, 170, April, 1940.

Maloff, I. G.: Electron Bombardment in Television Tubes, *Electronics*, **17**, 108, January, 1944.

——— and D. W. Epstein: Reflective Optics in Projection Television, *Electronics*, **17**, 98, December, 1944.

Markus, J.: New York–Boston Microwave Television Relay, *Electronics*, **21**, 114, January, 1948.

Mautner, R. S., and O. H. Schade: Television High Voltage Power Supplies, *RCA Rev.*, **8**, 43, March, 1947.

RMA Staff: Television Resolution Chart, *Electronics*, **20**, 123, December, 1947.

Roe, J. H.: New Television Field Pick-up Equipment Employing the Image Orthicon, *Proc. IRE*, **35**, 1532, December, 1947.

Schade, O. H.: Magnetic Deflection Circuits for Cathode-ray Tubes, *RCA Rev.*, **8**, 506, September, 1947.

———: Radio-frequency-operated High-voltage Supplies for Cathode-ray Tubes, *Proc. IRE*, **31**, 158, April, 1943.

Schlesinger, K.: Magnetic Deflection of Kinescopes, *Proc. IRE*, **35**, 813, August, 1947.

Seeley and Kimball: Transmission Lines as Coupling Elements in Television Equipment, *RCA Rev.*, **3** (4), 418, April, 1939.

Strieby, M. E.: Coaxial Cable System for Television Transmission, *Bell System Tech. J.*, **17**, 438, July, 1938.

——— and Weis: Television Transmission, *Proc. IRE*, **29**, 371, July, 1941

Television Transmitters, *Electronics*, **12** (3), 26, March, 1939.

Wendt and Fredendall: Automatic Frequency and Phase Control of Synchronization in Television Receivers, *Proc. IRE*, **31**, 7, January, 1943.

Wentworth, J. W.: Color Test Techniques for TV Broadcasters, *Electronics*, November, 1954.

Wright, A.: Television Receivers, *RCA Rev.*, **8**, 5, March, 1947.

Zworykin, V. K.: The Iconoscope, a Modern Version of the Electric Eye, *Proc. IRE*, **22**, 16, January, 1934.

———: Television, *J. Franklin Inst.*, **217**, 1, January, 1934.

CHAPTER 23

FACSIMILE

By A. G. Cooley[1]

1. Definitions. For the sake of convenience the following definitions of the terms used in discussions of facsimile are given. The definitions of many more terms will be found in the Institute of Radio Engineers' publications.[2]

Facsimile is the process or the result of the process by which fixed graphic material (subject copy), including pictures or images, is scanned and the information converted into electric signal waves used either locally or remotely to produce in record form a likeness (facsimile) of the subject copy.

Photofacsimile systems reproduce the subject copy on photographic papers or films.

Direct recording produces an image on a record sheet which is visible immediately upon recording, *i.e.*, without photographic developing or processing. Some direct-recording systems are intended only for black-and-white copy such as printed or typed matter, line drawings, etc., rather than photographs or full-tone pictures with areas of varying density. Other types are designed to give a reasonable approximation of photographic quality as in the electrolytic recorders. The end product of the recording process may, by the use of suitable recording material, be a facsimile duplicating master such as a hectograph master, mimeograph stencil, or printing plate.

Elemental area is any segment of a scanning line of the subject copy the dimension of which along the scanning line is exactly equal to the nominal line width. This has sometimes been called the scanning spot, although it is not necessarily the same. The terms "bit" and "bit rate" have also been used. In this discussion "element" and "element rate," respectively, will be used as being more directly applicable.

Picture Frequencies. The frequencies which result solely from scanning subject copy.

Note: This does not include frequencies which are part of a modulated carrier signal.

Base Band. In a carrier (or subcarrier) wire or radio-transmission system, the band of frequencies occupied by the signal before it modulates the carrier (or subcarrier) frequency to form the transmitted line or radio signal.

Note: The signal in the base band is usually distinguished from the line or radio signal by ranging over distinctly lower frequencies, which at the lower end relatively approach or may include d.c. (zero frequency). In the case of a facsimile signal before modulation on a subcarrier, the base band includes d.c.

Scanning. The process of analyzing successively in a predetermined pattern, the densities of the subject copy according to the elements.

2. Basic Principles. In a facsimile system the subject copy for transmission is scanned by a light beam, a line at a time, until the entire copy has been seen by the scanning beam. The transmitted or reflected light beam, modulated by the copy, is optically confined so as to pass on to a phototube at any instant the light from an elemental area, generally one one-hundredth inch square or less. The discrete area is referred to as an elemental area, or element for short. In the modulating system

[1] Times Facsimile Corp.
[2] IRE Standards on Facsimile: Definitions of Terms, 1956, *Proc. IRE*, **44** (6), 776, June, 1956.

of the scanner the light from the elemental areas usually controls the amplitude of the transmitted carrier signal. It is customary to use a carrier signal of such a frequency that each cycle of the signal represents one elemental area.

The amplitude of each carrier cycle represents the intensity of the reflected or transmitted light from the elemental area being seen by the phototube at that time.

The signals are transmitted to the facsimile receiver, which records each elemental area in a light or dark shade in accordance with the shading of the corresponding elemental area of the subject copy. The recording may be done by a photographic

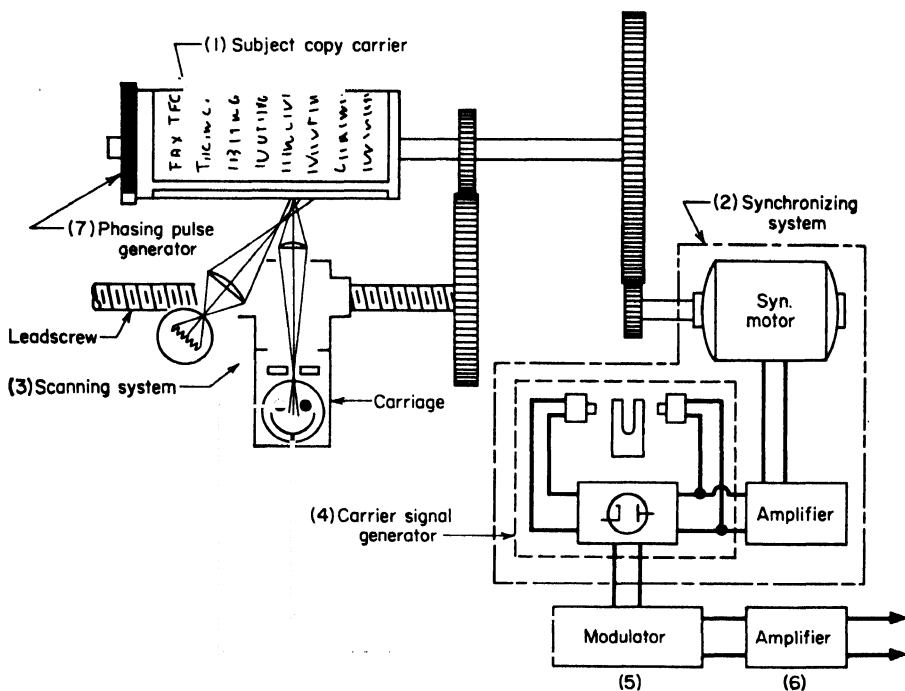



Fig. 1. Schematic of a basic facsimile transmitter.

process or a direct-recording process by means of electric current, heat, impact pressure, etc.

The transmitted signals are reproduced on the record sheet in a position corresponding to that of the transmitted elemental area.

3. Transmitter Components. A facsimile transmitter in an elementary form (Fig. 1) can be considered to be made up of the following basic units:

1. A subject copy carrier such as a revolving drum.
2. A synchronizing system for maintaining a constant predetermined scanning speed for synchronizing the machine with the recording speed of the recorder.
3. A scanning system including a light source for producing a beam of light which is reflected from the subject copy onto a light-sensitive device (phototube).
4. A carrier-signal generator.
5. A modulator in which the carrier signal is modulated by the output of the phototube.
6. An amplifier to increase the output of the modulator to produce a signal of sufficient level to transmit over a line or to control a radio transmitter.
7. A phasing pulse generator to indicate the start of a scanning line.

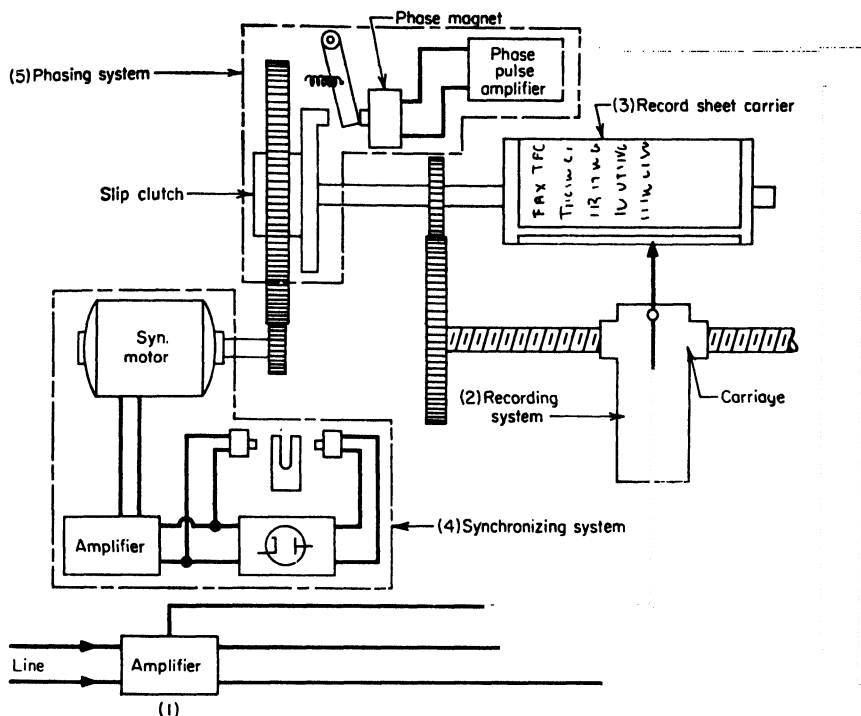



FIG. 2. Schematic of a basic facsimile receiver.

4. Recorder Components. A facsimile receiver in an elementary form (Fig. 2) can be considered to be made up of the following basic units:

1. An amplifier to produce sufficient power under the control of the received signal to operate a recording system.
2. A recording system such as a marking stylus that will burn off the white top coating of a conducting black paper.
3. A record-sheet carrier such as a revolving drum.
4. A synchronizing system for maintaining a constant recording speed corresponding to the scanning speed of the transmitter.
5. A phasing system to place the start of a received scanning line at the left-hand edge of the record sheet.

SUBJECT COPY AND RECORD SHEET CARRIERS

5. Drum-type Carrier. Cylinders or drums are most commonly used as subject-copy carriers. The rotation of the cylinder produces the relative movement in one dimension between the subject copy and the scanning or recording system with good dimensional accuracy and with a simple mechanical system.

Relative movement in the second dimension (normal to the direction of scanning) is produced by moving the scanning or recording system along the axis of the drum or moving the drum along its axis. The first method is illustrated in Fig. 1.

The principle of moving the drum along the axis is shown in Fig. 3. The drum shaft is threaded on one end with the number of threads per inch equal to the required scanning lines per inch. A nut engages the threads and is kept from turning by an extended arm that rides in a track. The nut pushes the drum as the shaft rotates.

Rotational motion is imparted to the drum through the shaft keyway and key attached to the drum.

"Drunken" threads or variations in angular velocity of the shaft will not produce a line-feed pattern or grouping of the recorded lines in the copy. This arrangement permits rigid mounting of the scanning or recording systems; therefore, no flexible connections are required. More space is required for the drum travel than with the moving carriage arrangement of Fig. 1.

Copy-holding Methods. The subject copy is held on the drum with various forms of clamps. One arrangement used extensively employs a "garter spring." The garter spring is rolled along the length of the drum just ahead of the scanned line (Fig. 3).

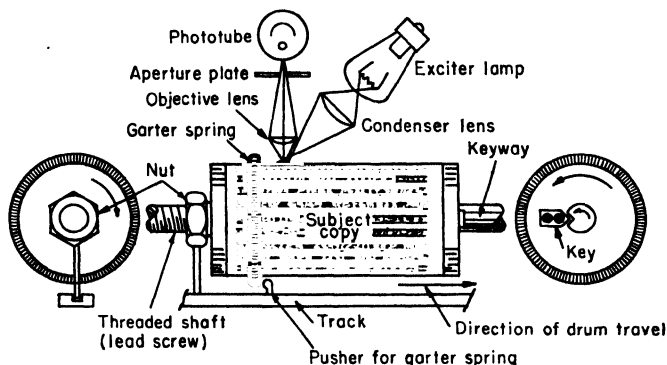



FIG. 3. Rotational and longitudinal motion imparted to drum through lead screw. Subject copy is fastened to drum by garter spring.

Some machines clamp only the leading edge of the copy and keep the trailing edge against the drum with cowlings or fingers. The copy is sometimes held against the drum with a transparent plastic sheet. There is no noticeable degradation in the recorded copy caused by the light beams passing through the transparent sheet in scanning the subject copy.

One system[1] makes use of centrifugal force to hold the copy against the inside of a transparent drum which revolves at 1,800 rpm. In another arrangement,[2] the copy is pressed face down against the outside of a nonrotating transparent drum. The scanning beams are rotated inside the drum. If two complete beam systems 180 deg apart are used, it is only necessary to wrap the copy halfway around the drum as the second beam system will start the scan of a line just after the first completes a scan.

Another arrangement[3] rolls the subject copy into a cylinder which is supported by an outside wrapper. A rotating mirror rotates the scanning beam inside the cylinder. The exciter lamp and phototube are external to the cylinder (Fig. 4). The diameter of the cylinder is about 1½ in. A similar arrangement is used for recording: the record sheet is rolled into a cylinder and the recording is done with an internal rotating stylus.

When a transparent photographic film is the subject copy, a transparent drum is employed so the scanning beam may be transmitted through the film.

6. Flat-bed Scanning and Recording. The term flat bed is applied to the method wherein the subject-copy and record-sheet carriers feed slowly in one direction while

[1] Cofer, T. F., Design of a Commercial Facsimile System, *Western Union Tech. Rev.*, **8**, 3, 105, July, 1954.

[2] Gillespie, C. N., "Automatic Facsimile Message Scanning System," U.S. Patent 2443953. Artzt, Maurice, Facsimile Transmission and Reception, "Radio Facsimile," Vol. I, p. 152, RCA Institutes Technical Press.

[3] Zabriskie, Douglas M., Ticketfax Transmitters and Recorders, *Western Union Tech. Rev.*, **9** (3), 93, July, 1955.

the scanning or recording across the sheet is accomplished by moving the light beams or recorder elements. The term "flat bed" is a misnomer as applied to some arrangements. The subject copy or the web (a continuous sheet) may be transported through the system by a partial wrap around a feed roll. The scanning or recording may be made across the part in contact with the roll. The record sheet is generally supplied

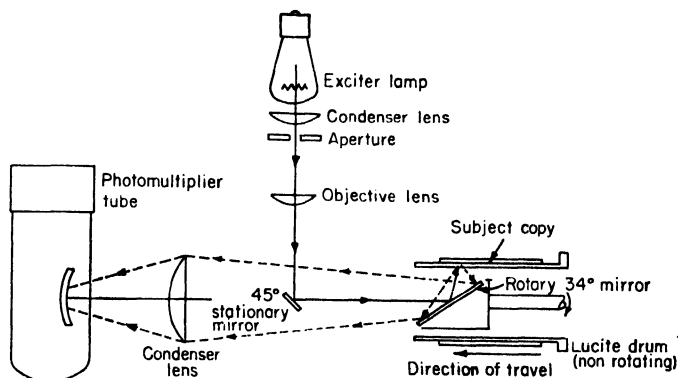



FIG. 4. Inside scanning system for nonrotating transparent drum. Western Union Ticketfax transmitter.

in the form of a roll or web. Some arrangements pull the web with gripper rollers close to the edge, and in some cases they push the web through the recording system.

SYNCHRONIZING SYSTEMS

The relative movement between the transmitter scanning beam and the subject copy must be synchronized with the relative movement between the recorder element and the record sheet. In a drum system this is accomplished by rotating the two drums at the same speed.

7. Requirements. Exact synchronization of the movement along the line of scanning is most important. Synchronization normal to the line is not so important. When transmitting message copy, a variation of 5 or 10 per cent in scanning normal to the line may not be objectionable.

Not only must synchronization of scanning and recording along the line of scanning be exact on a long-period basis but there must be no short-period variations as may be caused by variations in the angular velocity of the driving motor or nonuniformity in the mechanical coupling between the motor and scanning system. A variation in the mechanical load may reflect back into the motor, causing its angular position to advance or retard.

Short-period variations in synchronism produce a raggedness in the received copy known as "jitter." The effect is minimized if all errors in the system repeat each scan in exactly the same place. Since gear eccentricity is the most common source of jitter, use should be made of repetitive gearing in which each gear makes one or more full revolutions for each scanning cycle.

Photofacsimile recording systems are extremely sensitive to slight speed variations, which show up as density changes in the recorded copy. Density variations which are too small to be detected by sensitive densitometers may be very apparent on visual examination. Such density variations are usually caused by nonuniform velocity in the drive as the gear teeth change their contact point. The transmitter may have such speed variations without having an effect on the recorded copy.

8. Motors. Synchronous motors operating on the power line are frequently used to drive the scanning and recording mechanisms when the transmitter and

receiver are supplied from the same power source or different power sources which are synchronized.

A reluctance type of synchronous motor is an induction motor having a slotted rotor. The induction-motor action serves to start and drive the rotor close to synchronous speed. Slots in the rotor equal to the number of stator poles form a salient-pole rotor which will advance one pole for each pulse of the a-c driving power which pulls the rotor into synchronous speed after starting.

Another type is known as the "hysteresis" motor. Flux from the stator poles magnetizes the rotor with corresponding poles. The rotor is made of a solid steel alloy of the cobalt type and therefore has a strong coercive force that retains the magnetized poles. The rotating field of the stator pulls the rotor with it without slip after synchronous speed is reached.

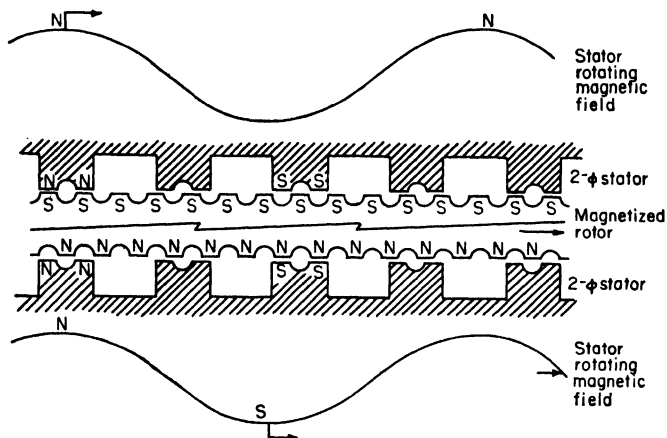



FIG. 5. Slow-speed 60-cycle synchronous motor. For each cycle of the magnetic field the rotor will advance one tooth.

The slotted-rotor induction motors and hysteresis motors operate at relatively high speeds, that is, 600 to 3,600 rpm if the power-line frequency is 60 cps.

A third type of self-starting power-line-frequency synchronous motor[1] manufactured by General Electric Company is designed for slow speed. Two of the common speeds are 75 and 100 rpm. This motor is a combination induction motor and phonic-wheel motor. The stator is wound as a two-phase induction motor. The rotor is split in the center, with one end magnetized south and the other north by a permanent magnet. Both the rotor and stator are notched to form a large number of salient poles. The line-up of the poles is such that, in maintaining the maximum pull between the rotating field of the stator and the magnetized rotor, the rotor will advance one pole or tooth each time the rotating field passes through one complete cycle. A motor operating on 60 cps will cause a 36-pole rotor to advance $\frac{1}{36}$ revolution for each cycle, or $\frac{60}{36}$ rps. There will be 100 rpm (Fig. 5).

If connected to a low-inertia load, these motors will start in one-half cycle. If the load inertia is high, mechanical decoupling between the load and motor is required. Ordinarily, a high load inertia is needed after starting to smooth out the power pulses in the motor drive.

This low-speed motor requires no more electrical power for starting than for running. The notched-rotor synchronous motor and the hysteresis motor demand starting power many times that of running power. This is no problem when operating from

[1] Staak, J. H., New Inductor Motor Has Low Speed, Self-start and High Torque, *Elec. Mfg.*, **39** (6), 115, June, 1947.

power lines. Facsimile systems intended for operation independently of power-line synchronization make use of a very stable local oscillator to control the frequency of the power applied to the synchronous motor. The motor-drive circuit includes vacuum tubes, gas tubes, or transistors. Rather than supplying motors requiring high starting power through the amplifier circuit, power from the power distribution line is connected to the motor momentarily.

High-frequency Motors. Synchronous motors of the variable-reluctance Lacour or phonic-wheel type manufactured by Times Facsimile Corporation[1] are employed when the frequency of the electrical power supplied to the motor is high compared with the required shaft speed (Fig. 6). These motors can be made to run efficiently on frequencies up to 3,600 cps and will drive light loads on frequencies as high as 6,000 cps. High-frequency operation eliminates or reduces the frequency-dividing requirements between the oscillator and the amplifier. Another advantage is that variations from no load to full load in the mechanical system produce only a very small angular displacement of the rotor, which has a large number of salient poles.

FIG. 6. Two-phase phonic-wheel synchronous motor.

Load Problems. In some cases where there is a problem of supplying sufficient power from the synchronous motor, an auxiliary motor is used to supply slightly more or slightly less than the average mechanical requirements.

The problem of large loads is handled in other systems by controlling a variable-speed motor with a phase-sensing device (Fig. 7). The principle is to generate a frequency directly proportional to the speed of the mechanical system and to match the frequency and phase with those of a stable oscillator.[2] A phase displacement

FIG. 7. Phase-error-sensing system for synchronizing.

between the two will cause the variable-speed motor to advance or retard sufficiently to correct for any phase displacement. Sometimes an eddy-current brake is employed to control the speed.[3] The braking power can be as much as ten times the electrical power applied to the brake. Such arrangements may be classed as servo systems involving error detectors and correctors. They are especially helpful in any system where there is a tendency for the load to reflect back into the motor.

The jitter produced by a load variation reflecting back to the motor is ordinarily a function of the ratio of the number of scanned elemental areas per second divided by the number of power cycles per second. For example, if two receivers operating at the same linear scanning speed in inches per second have a 60-cps motor in one and a 240-cps motor in the other, a load variation in the 240-cps motor set will produce

[1] Synchronous Motor Model MS, 2 Phase, Times Facsimile Corp. Service Bulletin, **9** (6), 33, October–November, 1955.

[2] Turner, F. T., L. G. Pollard, and C. R. Deibert, "Electric Motor Speed Regulation," U.S. Patent 2715202.

[3] Morgan, A. R., and M. Artzt, Color Television Recording, Part III, The Tape Transport Mechanism, *RCA Rev.*, **17** (3), 350, September, 1956.

one-fourth the jitter of the 60-cps set. This rule of thumb, however, does not take into account the motor "stiffness" or the kinetic energy of the rotors. The latter will vary as the square of the speed and is helpful in reducing effects of momentary load variations.

9. Amplifiers. Motor amplifiers, except those used for the phonic-wheel motors, may be of the conventional type using either vacuum or gas tubes in the output stage. Phonic-wheel motors which advance one pole per cycle require pulsating d-c power. The motor efficiency is best when there is no reversal of magnetism in the poles. To obtain this, plate current is passed directly through the motor coils. If the motor has high-impedance windings the required ampere-turns may be obtained with low plate current. To match the high motor impedance with that of the tube, an impedance-matching network is sometimes used.[1]

10. Synchronizing Oscillators. Oscillators for controlling the speed of the synchronous motors driving the facsimile equipment must have an exceptionally high degree of accuracy to prevent skew in the recorded copy. The degree of accuracy depends upon the number of elemental areas in the subject copy and the tolerable skew. News-photo machines transmitting 8- by 10-in. copy and scanning 96 lines per inch will produce satisfactory copy if the oscillators at the transmitter and receiver are operating within 1 part in 100,000 of the assigned frequency. Systems transmitting 12- by 18-in. weather maps require frequency accuracy of 1 part in 300,000. To transmit a full-sized newspaper page at 200 lines per in. the oscillators must operate within one part in 5×10^6 of the specified frequency to keep the skew in the received copy from appearing noticeable.

The frequency accuracy requirement in parts per million (ppm) is determined by

$$A = \frac{\Delta d \times 10^6}{2\Sigma L} \quad \text{ppm}$$

where Δd = tolerable skew in one transmission
ΣL = total length of all scanning lines in one transmission
ΣL is multiplied by 2 because allowance has to be made for the transmitter and receiver drifting in opposite directions from specified frequency.

11. Forks. Bimetal compensated tuning forks[2] operating at frequencies between 1,000 and 2,500 cps are commonly used to synchronize facsimile systems. The major portion of the bimetal or laminated fork is made of a nickel steel having a low temperature coefficient of expansion and a positive temperature coefficient of modulus of elasticity between 0 and +65°C. This is compensated by a lamination of carbon steel having a negative temperature coefficient of modulus of elasticity through that range. The bimetal forks ordinarily do not require temperature control or adjustments over periods of several years.

Specifications for an average bimetal tuning fork to be used in facsimile equipment are:

Stability:

1. ±3 ppm at room ambient temperature (no oven)
2. ±1 ppm after temperature equilibrium is attained ±5°C
3. Short-term drift rate less than 2 ppm per month during first 3 months of operation
4. Long-term drift less than 5 ppm per annum after first 3 months
5. Temperature compensated to less than ±0.2 ppm per °C from +10 to +65°C

Forks requiring a higher degree of stability are available by aging and selection. Stability of 1 part in 15 million is possible if the fork is maintained at a constant temperature and is aged for a year or more.

One type of fork is vacuum-sealed in a stainless-steel tube. The drive and pickup coils are external (Fig. 8). Final factory adjustment is made after the fork holds frequency over a wide temperature range. Service adjustments to take care of aging

[1] Cooley, A. G., "Synchronous Motor Circuit," U.S. Patent 2257158.
[2] "Fork Oscillators," Times Facsimile Corp. Service Bulletin, **10** (1), 3, January–February, 1956.

are made by varying the driving power. A 7:1 change in driving power will effect a frequency change of about 1 part in 20,000.

Tuning forks which are properly balanced will operate in any position with little effect on the frequency. Low-frequency (below 1,000 cps) tuning forks are not ordinarily used because the tines are difficult to balance and, being larger, are more sensitive to the direction of the force of gravity.

Since the frequency is affected by the drive, the output of the driving amplifier must be held constant. This is accomplished by the use of a limiter amplifier ahead of the output stage. Plate voltages are regulated within 2 per cent. It is not necessary to regulate the amplifier heaters (see Fig. 9).

Conventional frequency dividers are used when the fork frequency is a whole multiple of the motor-drive frequency. In facsimile equipment designed to operate at more than one speed, changes of speed may be made by changing the frequency of the power to the synchronous motor. This is accomplished by selecting a fork frequency which can be divided into the proper frequencies to give the required motor speeds. In some cases it is necessary to divide down to a low frequency and multiply up to a higher frequency. An oscillator which will both divide and multiply is shown in Fig. 10.[1] The wave shape across the resonant circuit

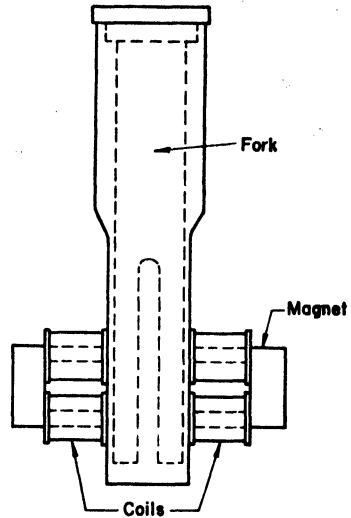



FIG. 8. Sealed tuning fork with drive and pickup coils.

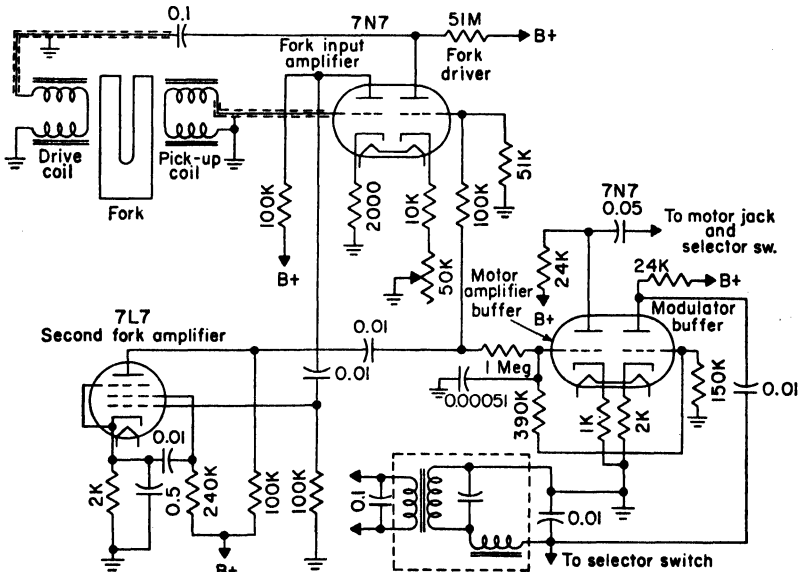



FIG. 9. Fork-oscillator unit.

[1] Sulzer, Peter G., Modified Locked-oscillator Frequency Dividers, *Proc. IRE*, **39** (12), 1535, December, 1951.

is approximately sine. If the resonant circuit is tuned close to the desired frequency, the circuit will lock over a fairly wide range of driving voltages. If locking over wide ranges of driving voltage and B plus voltages is desirable, a diode clipper referenced to the oscillator B plus must be used at the input. The oscillator is rich in harmonics, which makes it possible to operate at fractional divisions such as $2\frac{2}{3}$, by locking through control of a harmonic.

To simplify the problem of matching speeds between different systems, it has been suggested to manufacturers of equipment used in international services that the control frequencies be some multiple or submultiple of 300.

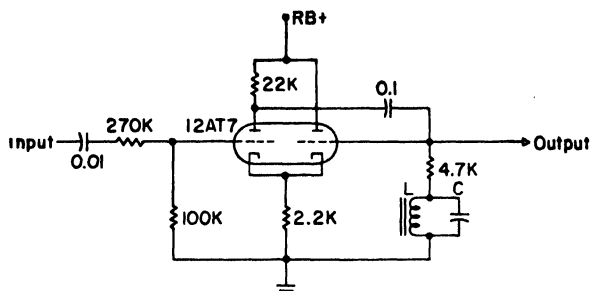



FIG. 10. Frequency-divider circuit. Feedback 0.1-μf capacitor is adjusted for output frequency which may be as low as one-fifth or one-sixth of the input.

Phonic-wheel motors can operate at the fork frequencies, thereby eliminating the requirement of frequency dividers. Hysteresis-type synchronous motors are now being manufactured which will operate at 1,000 cps. The shaft speed of these motors is exceptionally high. Muirhead hysteresis motors include a gear reducer in the motor frame.

TRANSMITTER-SCANNER OPTICAL SYSTEMS

12. Flood Projection. The most commonly used optical system for scanning subject copy is shown in Fig. 11. The subject copy is illuminated by a large spot of light. The image of the illuminated area is focused on a plate having an aperture to admit light to a photosensitive device. The light through the aperture represents the reflection from an elemental area.

Flood-projection systems are arranged so the specularly reflected light from a shiny or mirrorlike surface is not directed into the pickup or objective-lens system. The axis of the objective-lens system is usually perpendicular to the subject copy. The illuminating-system axis is usually displaced about 45 deg from the objective-lens axis.

A pair of planoconvex lenses are frequently used in the condensing system. A larger cone of light from the exciter lamp can be obtained with a single spherical-aspherical condensing lens. The aspherical side of the lens is shaped to correct the aberrations of a spherical system. This enables the use of a larger-diameter lens than would be practical with lenses having only spherical surfaces.

If the subject copy has creases or folds, shadow lines will appear in the recorded copy if a single beam of light is used to illuminate the scanned area. The shadow can be reduced by adding a second beam on the opposite side from the first.

Another way of reducing the shadow effect is to use an ellipsoidal mirror which directs the illuminating rays to the scanned area from all sides and at a steep angle (Fig. 12).

13. Spot Projection. A few facsimile systems employ the spot-projection arrangement shown in Fig. 13. The scanning spot does not exceed the size of the elemental

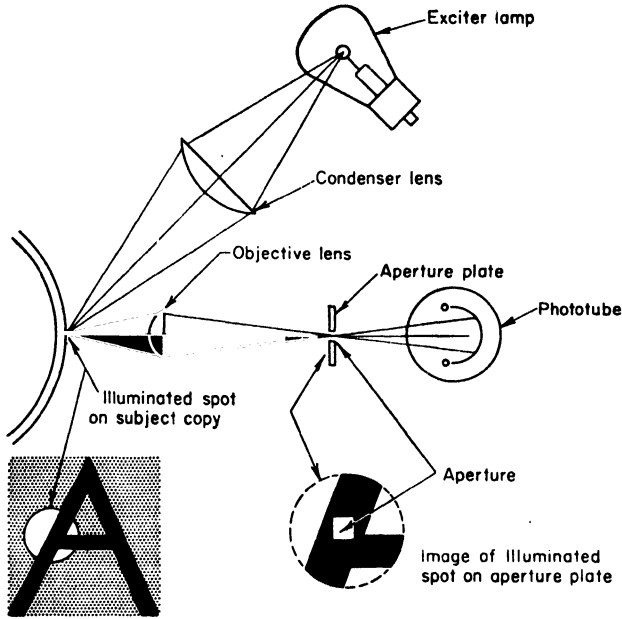



FIG. 11. Flood-projection scanning system.

area on the subject copy. Reflector or lens systems are often employed to aid the pickup of light reflected from the elemental area. Such a system is shielded from ambient light.

The objective lens must be corrected for spherical aberration. In most cases a pair of planoconvex lenses will suffice, but for improved results and elimination of the air space between lens elements it is customary to use a cemented lens having the corrections of a cemented doublet or a Hastings triplet.

Passage of scatter light through the aperture is kept to a minimum by stops in the lens barrel and blackening of the walls.

The aperture may be a square with one side equal to the width of the scanning line times the magnification of the system. If the image of the boundary between a black-and-white area is scanned, the phototube current will gradually change from minimum to maximum as the image of the boundary line crosses the aperture. Therefore, there is no sudden change of phototube

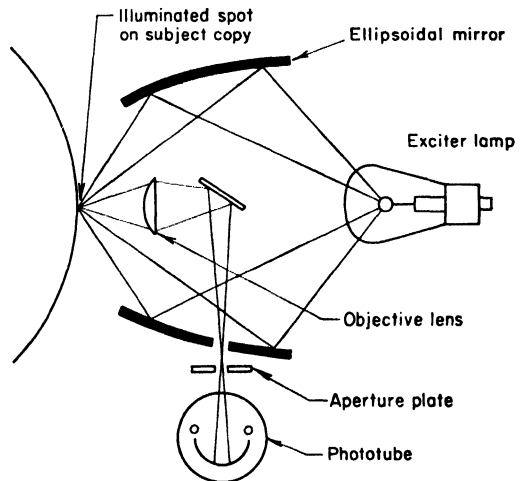



FIG. 12. Flood-projector scanning system with mirror instead of condenser lens.

current during a change from black to white in the copy. The failure to produce a sudden current change is known as *aperture distortion.*[1] In most systems the distortion in the transmitted signal will vary, depending upon the phase relation between the carrier pulse and the timing of the boundary-line passage over the aperture. Although aperture distortion is not an important factor in most systems, it can, as a refinement, be corrected or adjusted in the electric circuits or by changing the size or shape of the aperture. Sometimes the signal is overcorrected in the electric circuits to emphasize the border lines and assist in the transmission of thin lines which are less than the width of an elemental area. This is sometimes referred to as *unsharp masking.*

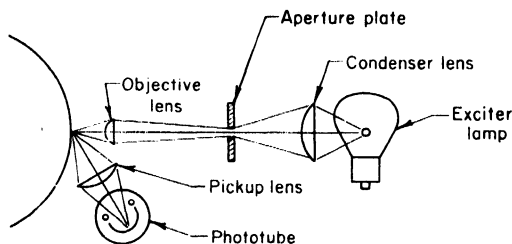



FIG. 13. Spot-projection scanning system.

Contrast. Scanning contrast of the optical system can be checked by imaging on the aperture a thin black line having a width which will just fill the aperture. The ratio of the line-image phototube current and white current should be at least 1:10, or 20 db.

14. Flying-spot Scanners. The spot-projection system is convenient to use in flat-bed flying-spot scanners. In mechanical systems the spot is swept by a moving mirror or rotating prism along the line to be scanned. In electronic systems operating at very high speeds the spot is formed on the face of a cathode-ray tube and swept across the face with conventional arrangements. It is then focused onto the subject being scanned (Fig. 14). Such a system was used by RCA in the "Ultrafax"[2] equipment, which demonstrated the transmission of 500,000 words per minute.

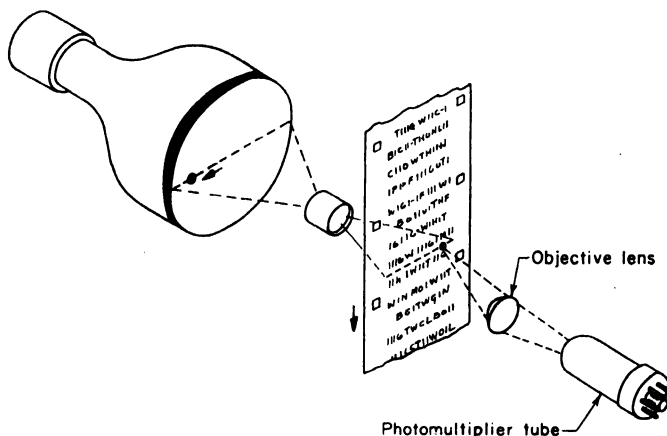



FIG. 14. Flat-bed flying-spot scanner of electronic type.

The interest in flying-spot scanners is based on the demand for a simple copy-loading system. Interest is centered around the four types shown in Figs. 15 through 18.

The *oscillating-mirror system*[3] is being developed by the Western Union Telegraph

[1] Artzt, Maurice, Facsimile Transmission and Reception, "Radio Facsimile," Vol. I, p. 153, RCA Institutes Technical Press.

[2] Bond, Donald S., and Vernon J. Duke, Ultrafax, *RCA Rev.*, **10** (1) 99, March, 1949.

[3] Buckingham, William D., A Flat-bed Facsimile Telegraph Transmitter, *Western Union Tech. Rev.*, **10** (3), 118, July, 1956.

Company (Fig. 15). It requires but little precision in the manufacture of mechanical parts because all mechanical errors repeat in the same position in each scanning line. The light source is a miniature tungsten-arc lamp developed for facsimile use. The light is emitted from a tungsten ball ten-thousandths of an inch in diameter. The mirror images the point of light on the copy to be scanned. The magnification of the optical system is 1. The mirror is oscillated by a cam which produces a fast return. A speed of 360 scans per minute is no problem. Tests have been made up to 1,800

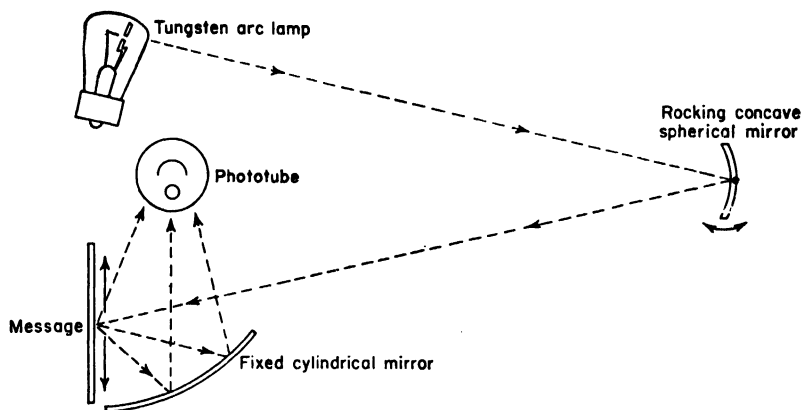



FIG. 15. Flat-bed flying-spot scanner using oscillating mirror.

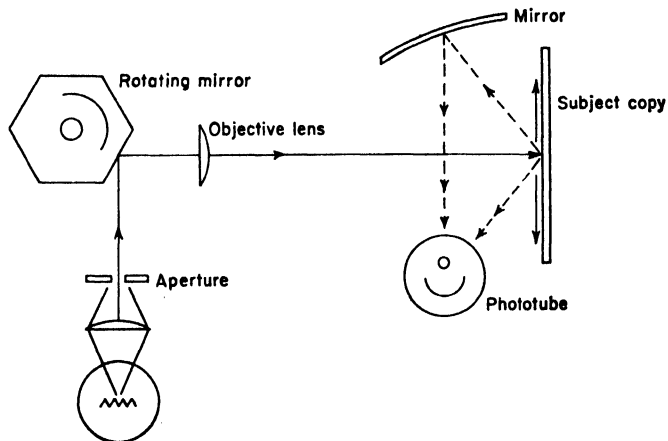



FIG. 16. Rotating-mirror type of flat-bed flying-spot scanner.

scans per minute. In the physical equipment the light beams are reflected back and forth between mirrors to reduce the space requirements of the equipment. This arrangement is referred to as "folded optics."

The *rotating-mirror system* of flying-spot scanning is being developed by the Signal Corps Engineering Laboratories and Times Facsimile Corporation (Fig. 16). It offers high optical efficiency and simple rotating mechanical parts. The precision requirements for the optics and mechanics are high.

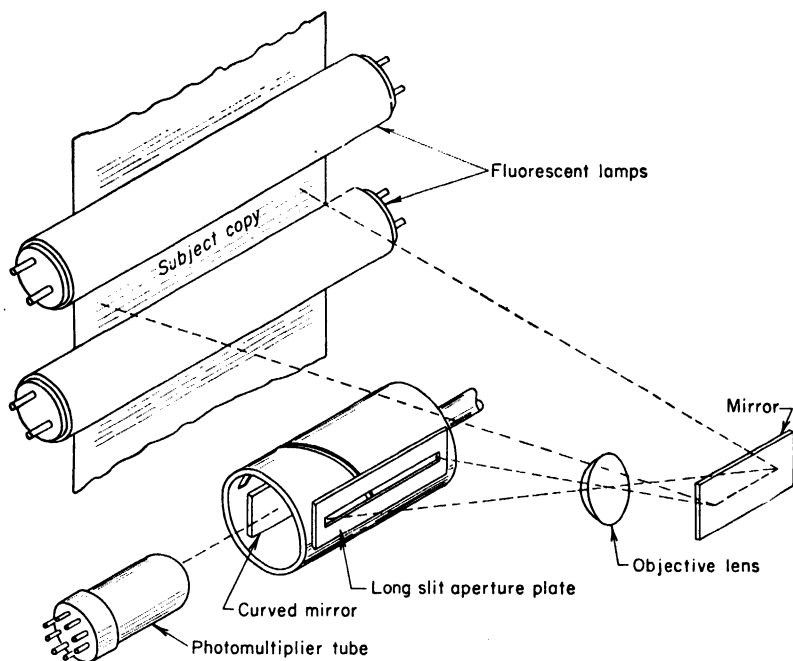



FIG. 17. Rotating-helix type of flat-bed flying-spot scanner.

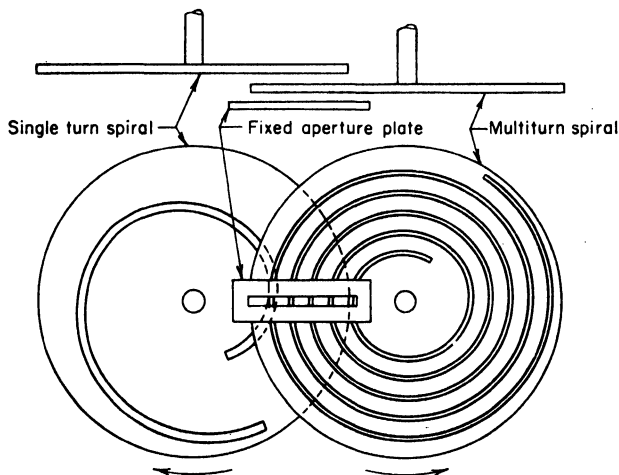



FIG. 18. Spiral-aperture system for flat-bed scanner.

Helix-type scanners[1] are manufactured by Alden Electronic and Impulse Recording Equipment Company. The aperture in a flood-projection-type optical system is the intersection of a long slit and a helical slit in a cylinder (see Fig. 17). A strip across the subject copy to be scanned is imaged on the aperture plate with the long

[1] Mathes, R. E., "Facsimile Reproducing Device," U.S. Patent 2167460.

slit. Light passing through the aperture is collected and reflected to a photomultiplier tube. The image on the slit is optically reduced so the physical dimensions of the rotating barrel with the helical slit will not be large. Space is also saved with folded optics.

Instead of the straight slot and helix, a system developed by the Hogan Laboratories, Inc.,[1] employs a slit in the form of a rotating spiral. The intersection of this slit and a long radial slit forms the aperture. In the physical make-up of the equipment, there are several turns of the spiral in the disk. A second revolving single-spiral disk masks off the unwanted spiral slits which are at the intersection with the stationary slit (see Fig. 18). The multiple-turn spiral is used to increase the angle of intersection with the stationary slit, thereby forming an aperture which is close to that of a square (see Fig. 18). Light passing through the aperture combination is directed to a photomultiplier tube. The image on the aperture system is optically reduced so large spiral disks are not required. Folded optics are also used in this arrangement.

LIGHT SOURCES AND TRANSDUCERS

15. Exciter Lamps. In the flood-projection systems low-voltage high-current filament-type lamps, such as those used for automobile headlights, are customary light sources. Short rugged filaments are desired as they do not sag and change position with use. When operated on a.c., the cooling during the zero portion of the cycle is less for heavy filaments than for the lower-current lamps, as shown in Fig. 19.

FIG. 19. A-c hum as a function of lamp current.

Spot-projection systems frequently employ a *crater lamp* of the R1130-B type[2] manufactured by Sylvania Electric Products, Inc. Rather than forming a spot by means of an aperture, the crater itself may be the spot source. While the lamp is designed for recording systems, it is useful as an exciter lamp in many transmitter scanning systems.

The exciter-lamp supply in a direct-recording facsimile system that is intended to record only black-and-white copy may be 60-cycle sine-wave a.c. if a lamp drawing 4 amp or more is used. The effect of cooling on lamps drawing lower currents may be reduced by supplying 60-cycle square-wave power. An approximation of square-wave power is developed in a saturable-core-type voltage regulator operating at its maximum rated voltage.[3] With this arrangement, the system may be used for transmitting photographs when operating with a 2-amp lamp.

The cooling effects of the lamp filament can be reduced by employing frequencies higher than 60 cps. The power supply in the military FX-1 and TXC-1 equipment supplies 1,800-cps power to the exciter lamp. The 1,800-cps signal is derived from the carrier frequency oscillator. The output of the power amplifier is accurately regulated by the system shown in Fig. 20.

An 1,800-cps signal is applied to a grid of a variable-gain tube. The output stage of the amplifier system drives a transformer having a 6-volt tap for the exciter lamp. This transformer has an additional output winding which feeds a rectifier. The rectified voltage

[1] Hogan, John V. L., and George M. Stamps, New Continuous-feed Facsimile Scanner, *Elec. Eng.*, **73** (7), 615, July, 1954.

[2] "Sylvania Glow Modulator Tube," Times Facsimile Corp. Service Bulletin, **8** (6), 29, July, 1954.

[3] Curtis, R. C., and S. J. Martin, "Exciter Lamp Operation," U.S. Patent 2496273.

is balanced against a reference voltage established by a 5651 reference-voltage tube. Any changes in balance are amplified and fed back to control one of the grids of the variable-gain input tube. If the output of the amplifier varies, the d.c. sent back to the input tube will change the gain and reestablish the lamp voltage to the correct value.

The control circuit consists of a rectifier which applies a negative voltage to the cathode of a multigrid tube. A constant voltage drop is inserted in the negative supply to the cathode by means of a 5651 regulator tube. One grid of the tube is connected to the supply ahead of the regulator tube so it is sensitive to voltage variations in the supply. Variations are therefore amplified and appear across the output resistor between plate and ground.

This system will hold the lamp-voltage variations to 1 per cent when the supply voltage to the power pack rapidly varies between 90 and 130 volts.

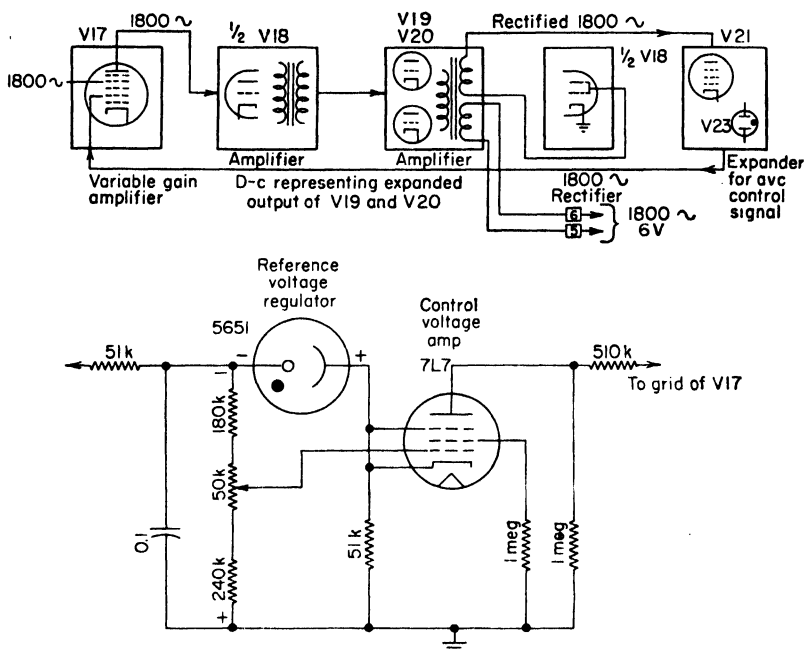



FIG. 20. (*Above*) Regulated 1,800-cycle exciter-lamp supply. (*Below*) Control-voltage amplifier.

Glow-modulator tubes used as exciter lamps are sometimes supplied with power at carrier frequency. Regulated d-c power is supplied to the crater lamp in other systems.

In the Telognosis system[1] for transmitting X-ray films of wide density range, the intensity of the light from a glow-modulator tube is controlled by feedback to produce an output signal of proper contrast. When scanning dense portions of the film the glow lamp current and intensity are increased above average, and they are decreased when scanning thin portions of the film, thus reducing the contrast range.

In transceivers, the crater lamp may serve as an exciter lamp while transmitting or a recorder lamp when recording.

16. Phototubes are used for interpreting the light from the elemental area in terms of electric current. Both gas and vacuum types are employed. The vacuum type

[1] McConnell, K. R., "The Transmission of X-ray Films by Facsimile," AIEE conference paper, June 1953 (available through Times Facsimile Corp.).

has a higher degree of stability but requires 10 to 20 db more amplification. The gas tube is generally unsatisfactory when a frequency response higher than 5,000 cps is required. When exposed to modulated light the sensitivity drops about 8 per cent between 1,000 and 5,000 cps.

The color response of the phototube is determined by the cathode material. This is in the form of a thin film on a metal support. A film of Co-O-Ag has its greatest sensitivity in the red end of the spectrum. Cathodes of this type are designated S-1. An S-4 coating of Cs-Sb has a more uniform response to the visual rays but its greatest sensitivity is in the blue region. Although facsimile recordings are usually black and white or intermediate shades, the subject copy is frequently in colors such as road maps. The S-4 coating is the more satisfactory for differentiating between colors, even though this is not apparent from Fig. 21.

Fig. 21. Spectral response of S-1 and S-4 coatings.

Since the S-4 coating is sensitive in the blue region, it is very susceptible to fluctuations in the exciter-lamp intensity if a filament-type lamp is used. Thus, any variations in filament temperature near the upper limit of its rating are exaggerated when phototubes having the S-4 coating are used.

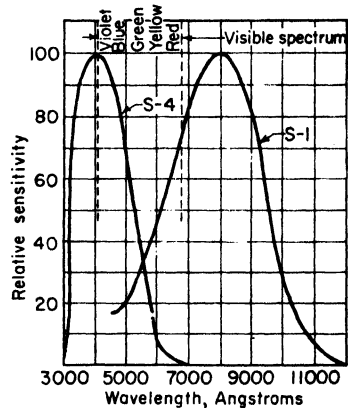
Multiplier-type phototubes are used when the light available from the elemental area is so low that the output from other types of phototubes is not sufficient to override microphonics and amplifier-tube emission noise. It is also helpful to have the high picture-frequency level of the multiplier tube to control modulating circuits.

Use of the photomultiplier tube eliminates the need for voltage amplifiers. The photomultiplier lacks the stability of a simple one-anode phototube and is very sensitive to variations in the dynode voltage.

Phototube sensitivity varies over a wide range. A phototube rated for 60 μa per lumen actually may have a sensitivity in the range of 40 to 90. This factor must be taken into account in designing amplifying systems. Furthermore, the sensitivity of the cathode is not always uniform over the entire area.

The required sensitivity of the phototube will depend upon the available light from the elemental area and the limitations of signal-to-noise ratio. The available light is so low in some flying-spot types of scanners that photomultiplier tubes must be used.

Other light-sensitive devices such as phototransistors, photovoltaic cells, and photoconductors[1] are not used in facsimile systems because of speed of response or temperature effects on sensitivity.

Microphonics and flexible leads are the two outstanding problems with optical systems mounted on a moving carriage such as that shown in Fig. 1. If a high-impedance phototube is used, it is often helpful to employ an amplifier having a low output impedance mounted adjacent to the phototube. A cascode stage offers certain advantages as a coupling element in meeting this requirement. Analysis of this amplifier shows that the wire joining the lower and upper stages operates at an impedance in the order of 1,000 ohms. This makes it convenient to separate the two halves—the lower half located close to the phototube and the upper half located at a distance. The circuit shown in Fig. 22 has a gain of about 800. The equivalent input noise is 4 μv. This is about one-fourth of the noise voltage generated by the

[1] Smollet, M., and J. A. Jenkins, The Lead Sulfide Photo-conductive Cell, *Electronic Eng.*, **28** (343), 373, September, 1956.

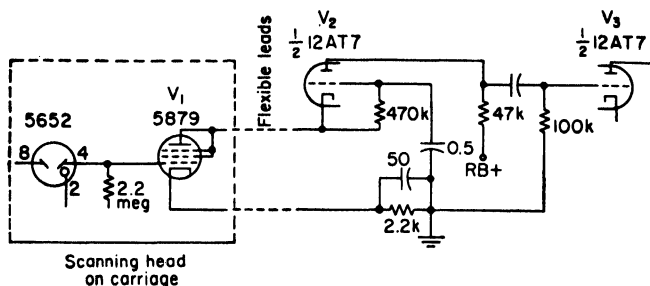



FIG. 22. Cascode coupling for phototube output.

2.2-megohm phototube load resistor. The equivalent input noise can be reduced to half by operating the filaments on d.c. The cascode depends on relatively high transconductance in the lower stage and high amplification factor in the upper stage. This makes it convenient to separate the two triodes, the lower (V_1) located close to the phototube and the upper (V_2) located at a distance. The circuit in Fig. 22 has a gain of about 800 from V_1 through V_3.

CARRIER-SIGNAL GENERATORS

17. Frequency. The lowest carrier-signal frequency that should be used is determined by the maximum number of elemental areas to be scanned per second and is limited by the maximum usable frequency band width of the communication channel. If single, lower, side-band transmission is to be employed, the carrier frequency can be about 40 per cent higher in the same available band width than it can for double side band.

Sharp cutoff filters for attenuating the upper side bands introduce a signal distortion so there must be a vestigial side band between the carrier frequency and the cutoff frequency of the filters or communication channel. In a communication channel having a cutoff frequency of 2,700 cps it is customary to employ a carrier of 2,400 cps. The general practice is to keep the carrier at least 10 per cent below the channel cutoff frequency.

The carrier frequency is sometimes set close to the cutoff frequency of the communication channel without any attempt to remove the upper side bands at the transmitter. These side bands are attenuated in the communication channel.

18. Synchronous Carriers. When transmitting to photorecorders which print the individual elemental areas that are distinguishable from each other, the carrier-signal generators are locked in with the synchronizing system. Unless a synchronous relation is attained, the elemental areas will print in a random pattern and produce a very noticeable effect. The effect is reduced or eliminated if the received signal is demodulated and smoothed out.

Lockin of the carrier with the synchronizing system is generally effected by controlling the synchronous motor from the same oscillator which supplies the carrier. The problem is minimized if the motor-drive frequency and carrier frequency are the same.

MODULATORS

19. Balanced Modulator. Modulators of the balanced type are normally used so as to reduce to a minimum the picture frequencies in the output of the modulator. Double balanced modulators such as the lattice type (ring type) prevent such frequencies from transmitting through a following amplifier. If the picture frequencies are present to a significant degree in the modulation products they cause severe distortion through heterodyne action with the carrier and the side bands. If a balanced modulator is not used, these frequencies can be reduced by a filter or the use of an

amplifier which discriminates against low frequencies, in some cases by the use of small coupling capacitors between stages.

If the picture frequencies (which might be described as "keying thump") are transmitted into the line, there is a good possibility that some of them will be attenuated less than the carrier with its modulation products. This will produce an etching effect in the recorded copy.

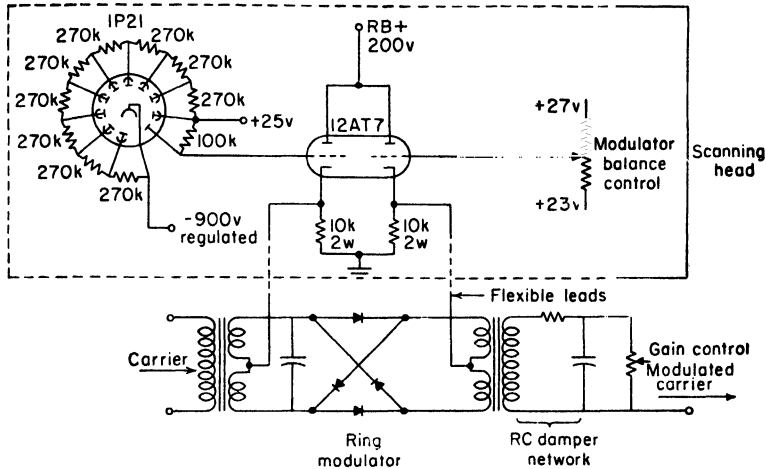



FIG. 23. Multiplier phototube and modulator.

A balanced modulator controlled by a photomultiplier tube is shown in one form by Fig. 23. This consists of a multiplier phototube, a pair of cathode followers, and a silicon-diode ring modulator. The phototube dynode-voltage divider is powered by a regulated negative voltage supply to stabilize the current amplification. The dynode-voltage divider is terminated in a stiff voltage source of about plus 25 volts. The two cathode-follower grids are also connected to this point—one through the phototube load resistor and the other through a potentiometer system for adjustment of a modulator balance. The two cathode followers form a balanced system which minimizes drifting of the modulator balance with changes in cathode temperature.[1] The phototube load resistor is not operated as part of the dynode bleeder. When light causes current to flow to the anode, its potential falls below that of dynode No. 9. This makes the sensitivity of the phototube fall off with increasing light. The voltage across the anode load resistor is approximately the log of the light intensity. Bifilar winding of the split coils of the modulator transformers assures balance of the leakage fluxes. The capacitor across the driver transformer should tune the transformer to resonance at the carrier frequency. The *RC* network damps ringing of the output transformer and partially rounds off the square wave which the modulator diodes deliver.

The cathode-follower grids are operated above ground so that large cathode resistors can be used to degenerate drifting further.

20. Phototube Modulator. A balanced modulator[2] which is used extensively is composed of a simple bridge circuit with the phototube in one arm. Light on the tube will balance or unbalance the bridge depending upon the modulation requirements.

[1] Hogan, J. V. L., and Frank A. Hester, "A Facsimile System Having Response Linear with Optical Density," A.I.E.E. conference paper, January, 1952 (available through Hogan Laboratories, Inc., New York, N.Y.).

[2] "The Photocell Bridge Circuit in TFC Transmitters," Times Facsimile Corp. Service Bulletin, **5** (1), 9, January–February, 1951. Shonnard, J. R., Facsimile Modulator Tube, *Electronics*, **21** (6), 82, June, 1948.

A simple, stable, balanced modulator is shown in Fig. 24. A carrier generator feeds the bridge approximately 1 volt peak (E). Dark balance occurs when $R_1 = R_2$ and $C_1 = C_2$. As light is admitted to the phototube, photoemission produces current through the load resistor R_L. The direction is controlled by the instantaneous phase of the carrier potential and the magnitude is proportional to the light flux. The resultant output of the bridge is a modulated carrier with upper and lower side bands. The picture-frequency band is eliminated when the phototube elements produce equal and opposite emission so that no d.c. appears in the load resistor R_L.

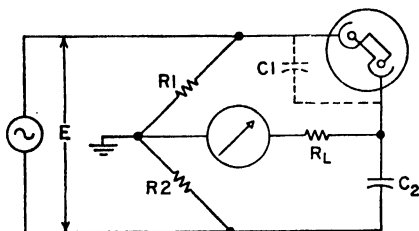



FIG. 24. Bridge modulator using double phototube.

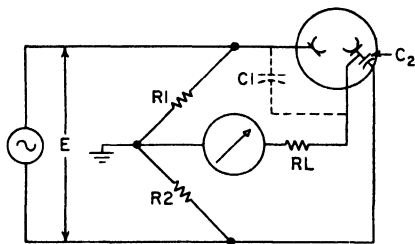



FIG. 25. Bridge modulator using RCA 5652 phototube.

The dark resistance of phototubes is very high, causing their circuits to be sensitive to leakage due to humidity. High-impedance wiring external to the tube also contributes to instability. To minimize these problems a special phototube was designed by RCA. This tube, designated 5652, is shown in the simplified circuit in Fig. 25.

The conventional anode is unnecessary because the cathode surface collects as well as emits electrons depending on the polarity of the applied signal. The active electrodes act alternatively as cathodes or anodes provided both surfaces are illuminated. Except for the external load all high-impedance wiring, including the neutralizing capacitor C_2, is incorporated within the tube envelope to provide maximum stability. Variation in photosensitivity of the individual cathodes is compensated by proper division of the light striking the two plates.

In *light-modulation* systems, the scanning beam is interrupted at carrier frequency. The interruptions may be effected by a light chopper or the source of light may be from a glow-modulator tube such as a Sylvania R1130B supplied with power at carrier frequency. The interrupter may be in the form of a slotted revolving disk or the vibrating shutter of a galvanometer. The beam is modulated by the light reflected from the elemental area. The picture frequencies must be removed.

SIGNAL AMPLIFIERS

21. Requirements. The recorder of a facsimile system is a form of recording oscillograph and thus shows any signal deterioration even though only a momentary transient is involved. A-c signal amplifiers must therefore be substantially free from microphonics, noise, hum, tendency for parasitics, etc. D-c amplifiers must be free of drift.

22. Gain Requirements. The maximum signal available from a phototube system in the average facsimile transmitter not employing a photomultiplier tube varies from 0.0002 to 0.1 volt. The amplifier output requirement is generally 0 dbm into 600 ohms.

A facsimile receiver supplied by a signal from a phone circuit will, in most cases, have an input level of −20 dbm (0.078 volt) to −10 dbm (0.25 volt). In most facsimile systems the power required to operate the recording element is less than 20 watts.

If the system is used for phototransmission, there must not be any sudden changes in gain which amount to ¼ db but long gradual level changes of even 1 db may not be noticed in the received copy. In black-and-white systems sudden changes of

1 db can be tolerated. These parameters apply to the combined amplifier and power-supply design.

Expanders and compressors are used in phototransmitters and receivers to effect tone correction. The expanders and compressors are designed to avoid any abrupt change in the slope of the curve showing gain vs. input-signal level. A circuit that will give the required correction for recording on Eastman Kodak transmission film, type C, is shown in Fig. 26. The minimum lamp current is 1.5 ma and the maximum 40 ma.

Amplifiers for phototransmitters are normally corrected to produce an output voltage which is linear with the density of the subject copy.

23. Frequency Requirements. In a double-side-band photofacsimile system the band-width requirement is usually in the range of one-half to one and a half times the carrier frequency. Allowing for Kendall effect this assumes the system is designed

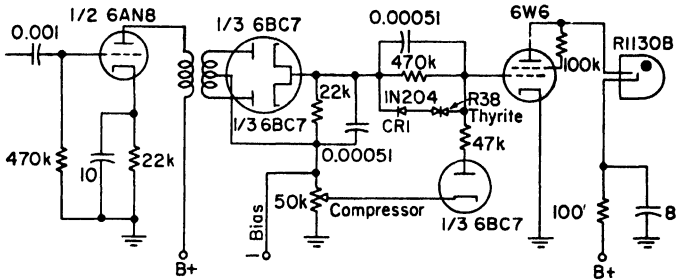



FIG. 26. Compandor circuit for recorder density correction.

to take advantage of the maximum usable keying frequency for the available carrier frequency.

In SSB systems the band-width requirement is generally from one-half the carrier frequency to the highest vestigial side-band frequency of importance (usually about 1.1 to 1.2 times the carrier frequency).

In the transmitter amplifier it is good practice to keep the frequency response over the band width employed within ±2 db of the gain at the carrier frequency. Satisfactory results are frequently obtained even if the gain drops off 6 db at the lower end, provided this drop is gradual. Such tolerances, however, do not apply to the receiver amplifier, if the facsimile signal into the receiver is a combination of audio-frequency shift (AFS) and a.m. This is frequently the case with signals delivered from a discriminator at the output of a radio receiver. In this case the signal contrast will be changed if the amplifier is not flat.

24. Phase Shift. The effects of phase shift in the amplifiers are usually not noticed in the received copy unless some type of filter is used to limit the amplifier band width. In these cases the phase shift must be linear with frequency to the closest practicable degree as any delay distortion in the outputs of the amplifiers will add directly to the delay distortions in the line. The total delay distortion for photofacsimile transmission of high quality should not exceed ±½ elemental area.

Some amplifiers are designed to invert the sense of the signal amplitude with respect to black or white. This is usually done by combining the signal from the phototube amplifier with another signal of opposite phase. The second signal will balance out a strong white signal from the phototube amplifier. In this type of signal inverter, the phase shift becomes very important. Phase shift may be purposely introduced to control the contrast by establishing the amplitude of the minimum signal at null. It is important that the phase shift remain constant in these systems since even a small drift with time or temperature would greatly affect the null amplitude and therefore the signal contrast.

25. Noise. It is generally considered that the peak noise in the system should be at least 10 db below the weakest signal which is to be recorded in a photofacsimile system. Many systems employ a 20-db contrast range in signal between black and white. This means the noise peaks should be at least 30 db below the maximum signal level. In black-and-white systems there is usually no objection to having the peak noise equal to the minimum transmitted signal unless the noise is of such a nature that peaks are very frequent. Good results are generally obtained even if the peak noise is above the minimum signal, and message copy is often readable with the usual size of type even when noise peaks greatly exceed the maximum signal.

Bias systems in amplifiers which operate class A present no problem. However, in an amplifier which does not maintain constant plate current the cathode bias resistors must not be by-passed with capacitors. Biasing capacitors cause a lag in the biasing voltage with respect to signal-level variations. Resistors in screen-grid circuits must not be by-passed if the average plate current is not constant.

26. Power Supplies. To prevent amplifier-gain fluctuations which may be caused by power-line variations, it is common to employ voltage-regulated power supplies. The combination of the amplifier and power supply is generally such that the output signal will not vary by more than $\pm\frac{1}{4}$ db when the power voltage is changed from 100 to 130 volts.[1] If the current drain on the power supply varies with the signal level, the time constant of the power supply often will give trouble by causing a change in gain unless the regulation is very good.

PHASING SYSTEMS

At the beginning of a transmission it is necessary to transmit a pulse indicating the start of a scanning line. This pulse is used to center the recorded image on the recording medium. The pulse may be a short break in maximum signal or it may be a short pulse of a maximum signal.

27. Pulse-generating Methods. Some systems transmit only a single phasing pulse at the beginning of a transmission. Others repeat the pulse for several seconds—in some cases up to a minute if operation is over a radio circuit. Transmitters are often equipped with timing devices to control the number of pulses transmitted prior to the start of copy transmission.

For generating the phasing signal the scanning beam in some systems either will scan white for the greater part of the scanning-line cycle and then a short black segment, or will scan black the greater part of the cycle, then a short white segment. In one system the scanning beam is blocked off during the greater part of the cycle and then permitted to strike a short white segment to produce the pulse.[2] This is done to eliminate false signals caused by scanning specks of dirt.

In some arrangements the phasing pulse is generated by a momentary closure of a switch or a sliding contact.

28. Recorder Phasing. Phasing is defined as the adjustment of the picture position along the recording line. In a broader sense phasing is the centering of the recorded image on the recording medium. In practice, the phasing pulse received from the transmitter determines the starting position of the recorded line. If the phasing is correct the recording of the line will start at the edge of the record sheet.

The phasing system shown in Fig. 2 is sometimes referred to as *stop-start* phasing. The receiver drum is stopped by the armature of the phase magnet engaging the clutch. A phasing pulse releases the clutch to start the recording at the edge of the record sheet. This system is commonly used where speeds are not great and the synchronous motor has sufficient power to operate against the stalled clutch.

The clutch friction must be great enough to pick up the mechanical load with very little slippage after it is released by the phasing pulse. The clutch-face materials

[1] "Voltage Regulator Circuit," Times Facsimile Corp. Service Bulletin, **10** (1), 1, January–February, 1956.

[2] Cooley, A. G., "Phasing Arrangement for Telefacsimile Machines," U.S. Patent 2246721.

are selected for a high ratio of static friction to sliding friction to prevent slippage during recording.

Some systems arrange to lock up the phase magnet immediately after the clutch is released for the first time. In other systems the phasing pulses are received over a period of several seconds so that the receiving operator can determine that phasing is correct. This is helpful when operating over radio circuits subject to random noise bursts which may cause false phasing unless there are such provisions for checking.

The phasing pulse is ordinarily demodulated and passed through a low-pass filter. This protects the receiver against false phasing on very short noise bursts.

In the Western Union Desk Fax[1] facsimile system which normally operates over physical circuits of short distances (8 miles or less), the phasing is accomplished by the transmission of a d-c pulse. This pulse is controlled by a commutator in the customer's transceiver. At the main office the pulse phases both the receiver and the transmitter with the customer's machine.

Some receivers driven by self-starting synchronous motors are brought into phase position by momentarily interrupting the power to the synchronous motor and permitting it to slip synchronism by one or more poles during each break.[2] The power interruption is repeated several times until the recording mechanism is in correct phase position. The indication is sometimes accomplished by having the receiver system develop a phasing pulse which is compared with the incoming phasing pulse. When the two are coincident, the system is in phase.

Instead of slipping the synchronous motor to bring the mechanism into phase, the motor is sometimes supplied with power of a frequency slightly higher or slightly lower than the synchronous frequency.[3] The receiver will then run slower or faster until the pulses are coincident.

If the speed of scanning is slow, considerable time is required to bring about phasing coincidence. If the recording speed is high, coincidence can normally be accomplished within 2 or 3 sec.

The coincidence system of phasing has the advantage over the stop-start system in that there is very little change of load on the synchronous motor during the phasing period.

Most of the advantages of both the stop-start phasing system and the *coincident phasing* system are realized in the *phase-corrector* phasing system. In such an arrangement the phasing error of the receiver is indicated upon the receipt of the first phasing pulse. A correcting mechanism operates until the indicated error is zero. In one form the correction is accomplished by rotating the frame of the synchronous motor.

DIRECT RECORDERS

Direct recorders approximating the schematic form shown in Fig. 1 are in wide use (over 30,000 in the United States in 1957). In addition to communications, similar facsimile principles are used for copying letters or forms and for producing duplicating masters.

29. Record Materials. The *dry electrosensitive* type of direct recorder uses a *record sheet* of the Timefax NDA[4] or Teledeltos[5] type. The surface of the sheet is a light gray, and although the resistance is high, current will pass through it to the lower-resistance undercoating or backing. Considerable heat develops under the recording stylus when a voltage is applied. This is due to the I^2R losses in the surface coating. The passage of signaling current causes disintegration of the top coating, exposing

[1] Ridings, G. H., and R. J. Wise, An Improved Desk-Fax Transceiver, *Western Union Tech. Rev.*, **6** (3), 92, July, 1952.

[2] Jelinek, Charles, Jr., Ticketfax Electronics, *Western Union Tech. Rev.*, **9** (3), 105, July, 1955. Artzt, M., "Broadcast Facsimile Transmission," U.S. Patent 2326740.

[3] Gillespie, C. N., "Phasing Apparatus," U.S. Patent 2496788.

[4] Times Facsimile Corp.

[5] Western Union Telegraph Co.

the black undercoating and thereby leaving a black mark. Writing speeds of over 1,000 in. per sec have been attained.

Teledeltos paper consists of a conducting carbon paper with a white pigment coating. Timefax paper has a coating of low-resistance carbon adhering to a white bond sheet. On top of the low-resistance layer is a high-resistance layer. This is coated with a modified zinc oxide pigment which is semiconducting. Variations of both types of recording papers are available which permit reproductions to be made by hectograph and litho-offset printing methods.

30. Recording Power Requirements. The amount of power required to mark direct-recording papers depends upon many factors, including the thickness and electrical characteristics of the electrosensitive surface coating, the stylus-to-paper speed, the stylus size and pressure, the frequency of the recording current, and the polarity when recording with d.c. A light pressure is desirable, and tungsten is preferred for the stylus tip to minimize wear, although steel is sometimes used. The

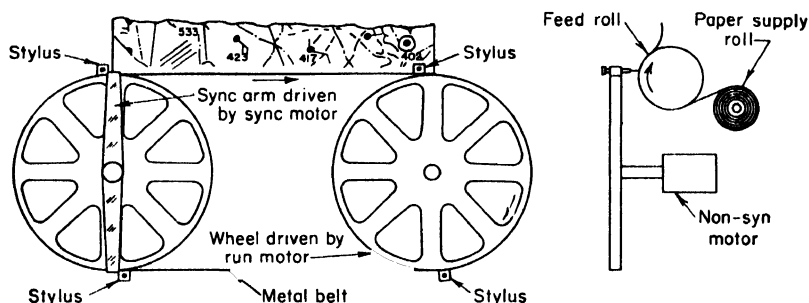



Fig. 27. Band drive for web-type continuous recorder.

power required for marking at different paper writing speeds varies directly as the speed. The voltage required increases only slightly with increase in speed. If d.c. is used with positive polarity on the stylus, about 7 watts is required for a maximum-density recording over a given area, with line spacing of 1⁄100 in., a paper writing speed of 50 in. per sec, and a stylus pressure of 7 g. The printing voltage for Timefax is about 225 volts for maximum density at 20 in. per sec and about 150 volts for Teledeltos. The maximum density is about 1.6 as measured on a Photovolt densitometer. This reading takes into account the surface characteristics of the paper and may differ from the readings obtained with glossy black, so far as the apparent density or "blackness" on visual inspection is concerned.

31. Web-type Recording. The trend in direct recorders is toward the continuous-web type as they do not require reloading after each transmission and will operate without attention. Many of the dry electrosensitive types are used in meteorological services, and there is a need for them in message-communications operations. The styluses are moved across the width of the web by a continuous band driven at synchronous speed. Three or four styluses are attached to the band so that, when one recording line is completed, the following stylus will start the next line.

The Western Union system drives the band and maintains its phase position by sprocket holes in the band.[1] The Times Facsimile system employs a nonsynchronous drive to one of the wheels on which the band is mounted (Fig. 27). The band is prevented from running above synchronous speed by engagement of lugs on the band with a synchronously driven arm rotating concentrically with one of the wheels.[2]

[1] Hallden, F. G., Three Stylus Facsimile Recorder for Concentrator Service, *Western Union Tech. Rev.*, **4**, 16, January, 1950.

[2] Cooley, A. G., and P. L. Grafstein, "A New 18-inch Wide Continuous Web Facsimile Recorder," Conference Paper CP55-774, 1955 (available through Times Facsimile Corp.). Cooley, A. G., and P. L. Grafstein, "Design Factors in Continuous Facsimile Recorders," A.I.E.E. conference paper, January, 1954 (available through Times Facsimile Corp.).

The Western Union system has the styluses attached rigidly to the band. The web along the line of scanning has a resilient backing. The backing for the web in the Times machine is rigid but the styluses are resiliently mounted on the band.

32. Pressure-sensitive Recorders. These are used for printing mailing-address labels.[1] Complete four-line labels having 25 characters per line are printed at the rate of 200 per minute. The marking is done by pressure applied to a wax-carbon paper. This requires two webs, the record sheet and the carbon paper.

A single-sheet pressure-sensitive arrangement has proved out in prototype machines and can be expected to reach commercial production. The paper is made up of a

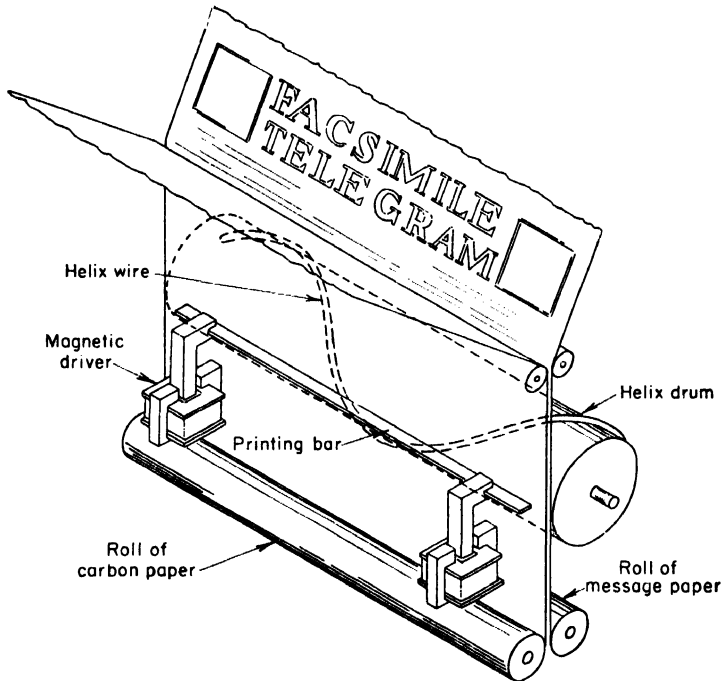



FIG. 28. Pressure-sensitive recorder, helix and bar type.

sheet of bond paper on which is a coating of black ink or dye. The dark coating is masked with a thin layer of wax or metallic soap. The metallic soap is preferable since the melting point is above 200°F. The masking layer contains air bubbles which produce a white appearance. Pressure from the marking stylus compresses the wax or metallic soap and makes it transparent, thus exposing the dark undercoating.

33. Helix and Bar Recorders. In this system (Fig. 28), sometimes called *lawn-mower* recording, the pressure-sensitive sheet is passed between a rotating helix and a bar.[2] When the bar is magnetically moved toward the helix a mark will be made on the record sheet along the line of intersection of the helix blade and the bar. The helix is driven at synchronous speed so the intersection point makes one complete sweep across the record sheet during each scanning cycle.

Magnetically driven recorders of the helix and bar type are limited in printing speed by the mass of the bar. Recorders having a bar length of 8 in. have found only limited use. The recorders for printing mailing labels referred to above have a

[1] "Tape Facsimile Addressing Machines," Times Facsimile Corp. Service Bulletin, 7 (9), 39, October, 1953.

[2] Young, C. J., "Picture Receiving Apparatus," U.S. Patent 1848862 (reissued as Re. 20152).

bar that is about 3/16 in. long and prints only one line of type. Four recording heads are used to print the four lines of the labels. Each head has a printing capacity of 7,000 elemental areas per second (keying frequency of 3,500 cps).

The U.S. Army Signal Engineering Laboratories have had success with a direct-recording system employing three magnetically operated styluses carried on a band in a continuous-web recorder. The unique magnetic system is shown in Fig. 29. This arrangement reduces the travel requirements of the armature to the amount the pressure-sensitive record sheet compresses under the pressure of the stylus.

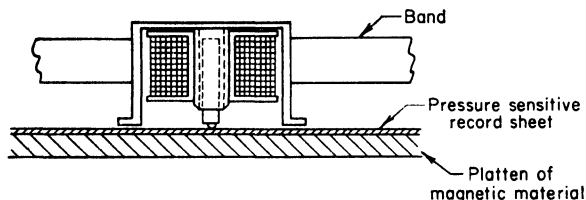



Fig. 29. Magnetically driven stylus for recording on pressure-sensitive record sheet.

34. Ink Recorders. Ink is used as the marking medium in a Siemens-Hell-Fax recorder of the drum type manufactured in Germany. The magnetically actuated stylus is a disk 1/8 in. in diameter and is referred to as the "writing wheel." The writing wheel is inked from a driven rubber ring, the surface of which is inked by an inking roller.

Ink has been used in helix and bar systems. The ink was applied to the helix through a transfer roll. A few of these machines are still in service, although the

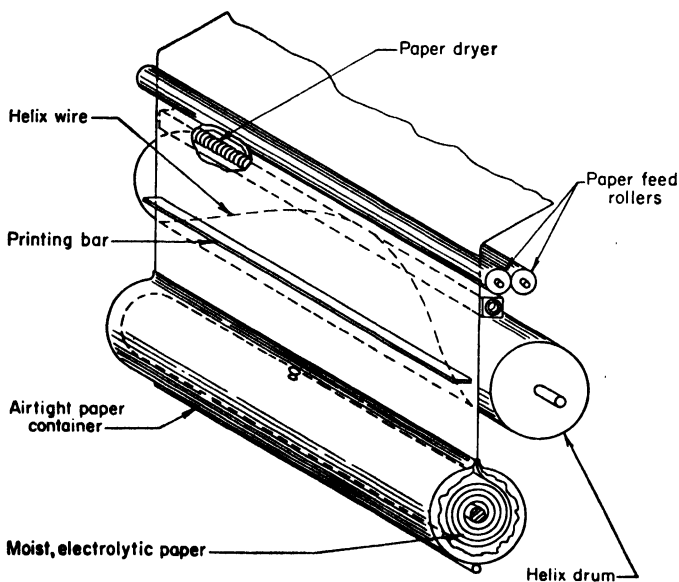



Fig. 30. Helix and bar type of electrolytic recorder.

Hell-Schreiber tape of facsimile equipment of this type was used quite extensively at one time in radio press services. Ninety-eight elements were required to make up one character. Only seven elements or bits are required for one character in a printing telegraph system; so the tape facsimile could not compete in this service. The tape facsimile is, however, used as a stand-by for operation when radio conditions are too poor for printing telegraphy.

35. Electrolytic recorders[1] are being used in news-photo services for newspapers and television stations. Some electrolytic recorders are used for recording weather maps and they are being developed for message-communications service. The translucent recording paper may be used as a master in light-sensitive duplicating machines such as Ozalid or Bruning.

Electrolytic recorders are generally of the helix and bar type shown in Fig. 30. The record sheet is moistened with an electrolyte. Current passes through the sheet at the crossover point of the bar and helix. In some systems the electrolyte colors upon the passage of current.[2] Paper developed by Hogan Laboratories gives a black recording by dissolving iron from the steel marking anode with an alkaline substance formed in the paper by the passage of current. The paper may contain sodium nitrate, pyrocatechin and water, for example, and the iron reacts with the pyrocatechin in the presence of the alkali to form a dark substance (color lake) at the recording point.

PHOTORECORDERS

Photofacsimile recorders find their greatest use in news-photo services. Radiophoto services that handle picture traffic use the photorecorders for black-and-white as well as continuous-tone subjects. Some use is being made of photofacsimile for transmitting newspaper copy for proofreading and for publication from the received film. Photofacsimile recordings have been made at speeds up to a million words a minute over very-wide-band channels by the Ultrafax® system. The Ultrafax recorder is similar to the scanner shown in Fig. 14 except for the photomultiplier tube. Negative film is transported in the place of the subject copy.

36. Glow-lamp Recording. Either film or a bromide paper is exposed in the photo-facsimile recorders by a light which varies in intensity at the carrier or modulation frequency. Most of the recorders manufactured today use the glow-modulator tube type R1130-B. The light output is proportional to the current through the lamp and the voltage across the lamp remains fairly constant even though the current may vary over a range of 50 to 1. A cross section of the glow-modulator tube R1130-B is shown in Fig. 31. The characteristics are shown by Fig. 32. The light forms along the walls of the crater in the zirconium cathode when a positive potential is applied to the anode. Looking into the crater, the light appears doughnut-shaped.

FIG. 31. Glow-tube modulator, Sylvania R1130-B.

A positive potential exists along the axis of the crater and the walls are at negative potential. This causes electrons emitted from the walls to be accelerated laterally to the center. They continue toward the opposite wall but are repelled by the potential minimum there. This oscillation occurs many times before the electrons are eventually drawn out of the crater to the anode, thereby providing an extra long path for the electrons.

The ions and the electrons carry the discharge current through the tube while the excited molecules of cathode vapor emit light when they return to their normal state. In effect, a vaporizing of the cathode material is the source of light. Light produced by the glow of the fill gas is negligible. The modulation characteristics are therefore a function not of the ionization and deionization time of the gas but of

[1] Hogan, J. V. L., H. C. Ressler, and W. H. Tribble, "Electrolytic Recording," U.S. Patent 2358839. Hogan, J. V. L., J. W. Smith, and A. H. Mones, "Electrolytic Recording for Facsimile," A.I.E.E. conference paper, February, 1955 (available through Hogan Laboratories, Inc., New York, N.Y.).

[2] U.S. Patents 2339267 and 2358839.

the rate of heating and cooling of the cathode vapor discharge. The frequency response is dependent upon the gas pressure which affects the cooling rate. The total light output goes down at a given current as the pressure of the gas filling increases because of the increased rate of heat conduction away from the crater. For modulating frequencies in the range of 50 kc, a gas pressure of 25 mm is used. The current may be as high as 150 ma for periods of several minutes. During the first 5 or 10 sec of operation of a high-pressure lamp, there is an increase in light output.

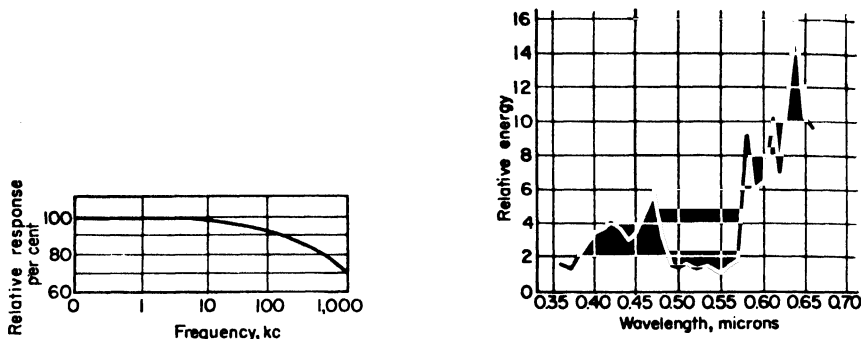



FIG. 32. Characteristics of the R1130-B glow tube.

Mica disks positioned above the opening in the anode help to collect material from the cathode which would otherwise be deposited on the window of the envelope, thereby cutting down the light.

A typical optical system for a glow-modulator tube is shown in Fig. 33. The objective lens focuses the image of the aperture on the record sheet. The highest optical efficiency is obtained when the aperture plate and field lens are close to the envelope of the glow tube. Under these conditions, however, the objective lens will image the lamp crater as well as the aperture on the record sheet, thus producing an undesirable spot for recording.

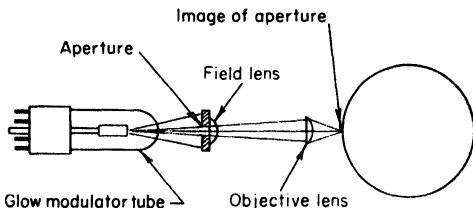



FIG. 33. Recorder optical system using glow-tube modulator.

One purpose of the field lens is to focus the image of the crater on the front surface of the objective lens. If this could be fully accomplished, the image of the aperture on the record sheet would not be blurred with an unwanted image of the crater. Because of the depth of the crater, a sharp image cannot be placed on the surface of the objective lens unless the spacing between the crater and aperture is great. This is because the anterior depth of focus of the objective lens is large. A diffuser between the crater and aperture will prevent the objective lens from seeing the crater, but this advantage is obtained at the cost of light efficiency, generally a 2 to 1 reduction.

If maximum optical efficiency is required, the crater can be focused on the record sheet with a single lens, omitting the aperture and field lens. With this arrangement, the image is round and not clearly defined. The recorded elemental areas are then not of uniform density, and they build up into recorded lines which vary in density across the lines.

It is common to use objective lenses having focal lengths of ½ to 1 in. The optical efficiency can be improved a small amount by using shorter-focal-length lenses, but

the posterior depth of focus is so short that any minor bulges in the record sheet or an irregularity of the record-sheet carrier causes an out-of-focus effect on the recorded copy. Devices for clamping the record sheet on the carrier (drum) protrude beyond the surface and determine the shortest focal length which may be used without risking mechanical interference between lens tube and record-sheet clamps.

37. Films. Eastman Kodak Transmission films type A and type C are commonly used in news-photo work. The *H* and *D* curves are shown in Fig. 34. Both type A and C can be developed, fixed, and dried rapidly. For high-speed recording Contrast Process Ortho[1] has proved satisfactory.

Enlarging paper such as Kodabromide F1[1] is suitable as the record sheet when positive prints are required.

FIG. 34. *H* and *D* curves for Eastman Kodak types A and C films.

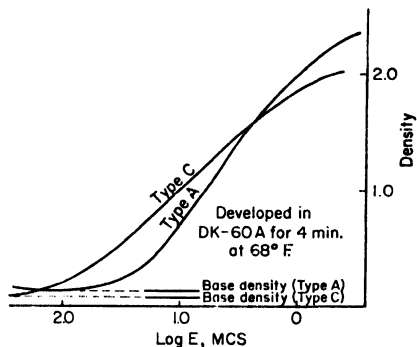
38. Positive and Negative Recording. This technique can be accomplished on the same machine without reversing the scanning direction with respect to the line advance by placing the film on the drum with the emulsion side down so the image will appear in the same left-to-right sense as camera negatives. Placing the emulsion side down also has a handling advantage in that fingerprints on the back of the film do not appear in the finished negative. The glow-modulator tubes produce sufficient light to expose the emulsion properly after passing through the antihalation backing.

If a *mechanical system* such as shown in Fig. 2 is used to move the recorder optical system along the length of the drum, careful consideration must be given to the design and construction. Irregularities produce an underlap or overlap of the recorded elemental areas. Irregular underlap and overlap produce undesirable patterns. This is called *grouping* when the variations develop patterns the full length of the scanning lines. If the recording is at the rate of 100 lines per inch, a variation in underlap or overlap of 0.0005 in. is usually acceptable, depending somewhat upon the sharpness of the recording spot and the rate of variation.

Very small variations in the lead-screw pitch will produce grouping unless the number of lead-screw threads per unit length is equal to the number of recording lines per unit length. Grouping is also caused by a cyclical variation of the lead screw along its axis as it rotates. To keep this type of camming to a minimum, the lead screw is usually end-loaded against a cone bearing or a single-ball thrust bearing. A slight amount of eccentricity of the lead-screw gears will cause grouping. In most cases the eccentricity of the pitch line must be under 0.001 in. The carriage must fit firmly to the ways so the optical system cannot vibrate in any direction. The carriage must slide without any variation in load that will reflect back to the gears and motor.

TELEPHONE CIRCUITS FOR FACSIMILE TRANSMISSION[2]

There are now over 114,000 miles of telephone circuits in the United States which are being used for facsimile. Telephone circuits that are suitable for voice communications are not necessarily satisfactory for facsimile. The problems of wire-line transmission must be considered.

39. Requirements. Facsimile transmission over telephone circuits does not place severe demands on the communication channel if the copy to be transmitted is essentially black and white, in which legibility is the primary criterion and appearance

[1] Eastman Kodak Co.
[2] By A. C. Mitchell, Times Facsimile Corporation.

only secondary. The requirements for quality of transmission are much greater than in the case of voice communications if consideration must be given to many degrees of middle tones and complete freedom from visible effects of level changes, noise, frequency patterns, transients, echoes, and delay distortion. This second and more demanding class of material always requires photographic or high-quality electrolytic processes to produce the received copy while the first makes use of conventional direct-recording methods.

For example, the relation between facsimile machine characteristics and transmission qualities of telephone lines can be determined as follows.

Assume the requirement is to transmit 8- by 10-in. subject copy at a resolution of 100 lines per inch along the scanning line and normal to the scanning line, and that the transmission time is to be 8 min. The required element rate (per second) would be

$$\frac{8 \times 10 \times 100 \times 100}{8 \times 60} = 1{,}665 \text{ elements per sec}$$

To transmit this material in 8 min, however, allowance has to be made for a small portion of the scanning operation used for blanking, or phasing, for instance, the clamp bar of a drum-type scanner. If we assume 8 per cent for this we get 1,665 + 135 = 1,800 elements per second as the rate needed to transmit the copy in 8 min.

The effective band for facsimile transmission would then be 900 cycles, and, to avoid overlap in the voice-frequency portions of the telephone lines between the base-band frequencies (0 to 900 cps) and the lower side band of a carrier system, the carrier might be placed at 1,800 cps if double-side-band transmission is involved. The two side bands would then extend from 900 to 2,700 cycles.

This band could be accommodated by almost any type of telephone-message circuit for widely different limiting lengths as far as frequency band width is concerned. (One exception would be an "E.B." split channel.) This does not, however, necessarily mean that an ordinary telephone-message circuit will be satisfactory in every respect. The reasons for this are explained below under the several types of unwanted phenomena.

40. Envelope-delay Distortion. This phenomenon can be ignored in voice transmission but is of great importance in facsimile. Envelope-delay distortion is a measure of the difference in propagation time of the frequencies making up the envelope of a transmitted signal element.

For high-quality transmission of original photographs, experience has shown that the delay distortion should not exceed ±½ elementary area (element). This limit expressed in microseconds would be inversely proportional to the element rate employed in transmission. For example, at 1,800 elements per second, used in the example above, the time of half an element would be

$$t = \frac{1{,}000{,}000}{2 \times 1{,}800} = 280\ \mu\text{sec}$$

The limit for envelope-delay distortion over the band of modulation frequencies from the carrier minus 900 cps to the carrier plus 900 cps should not exceed ±280 μsec for the most exacting type of facsimile transmission with photographic-recording processes. If instead of 1,800 elements the rate was 18,000, the permissible delay distortion would be only ±28 μsec.

In the case of black-and-white typed or printed material where legibility is the most important consideration and with existing direct-recording processes, ±500 μsec would be tolerable at an element rate of 1,800.

As for the severity of this trouble for the same lengths of circuits, telephone systems making use of carrier techniques usually have much less delay distortion than those which do not use such techniques. The delay distortion within the usual facsimile band of 900 to 2,700 cps for 100 miles of certain types of voice-frequency loaded-cable system might be as much as twenty times greater than one 100-mile *K*-carrier link. In carrier-frequency systems the distortion is mainly in the terminal equipments

so that it is roughly proportional to the number of links of carrier channel rather than the length of the circuit. In voice-frequency loaded-cable systems nearly all the distortion is in the loaded cable itself and is closely proportional to the length. See curves of Fig. 35.

Correction for delay distortion is usually made by installing the necessary number of sections and characteristics of passive-delay equalizer networks. Variable active-delay equalizers are also available.[1] In any case the delay equalizing is done by adding more delay at the frequencies having small value of delay to make the delay more nearly the same for all picture frequencies.

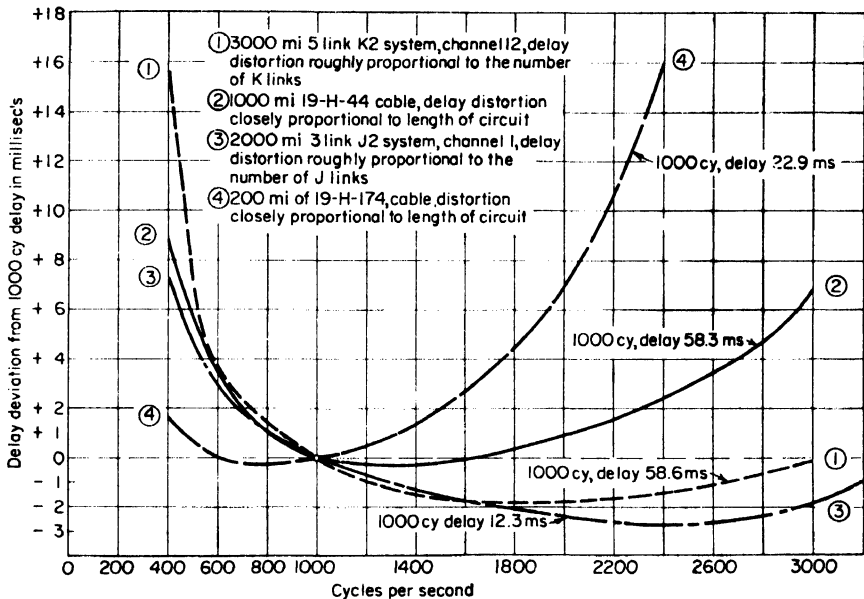



Fig. 35. Envelope delay of four representative telephone circuits.

41. Attenuation Distortion. Facsimile circuits should be held to a flat net loss-frequency characteristic between 1,000 and 2,800 cps within considerably narrower limits than are required for satisfactory telephone transmission. Facsimile introduces a further complication in that the carrier frequency usually employed is 1,800 cps or higher, whereas all telephone-circuit layouts are based on 1,000-cps attenuation and loss data. In circuits especially designed for facsimile transmission, special attenuation equalizers are installed in various sections of the circuit depending to some extent on its length.

The most serious attenuation problem is frequently in the local exchange area when the local channels sometimes are fairly long nonloaded cables of No. 19, 22, 24, or smaller gages where the attenuation at a picture carrier of, say, 2,400 cps may be many times as much as at 1,000 cps (see Table 1).

Another peculiarity of exchange-area cables is that, for telephone service, impedance irregularities in such cables usually are of little importance and therefore are prevalent. Two such irregularities in a loaded cable will react on each other to produce wiggles in the attenuation-frequency characteristic which, while not in themselves important, may produce associated peaks and valleys in the delay-frequency characteristic the results of which may be visible on photographically processed recordings.

[1] Cannon, W. D., "Delay Distortion Correction," AIEE Transactions Paper No. 56-107, January, 1956.

42. Level Changes. The most frequent cause of complaint in facsimile, particularly photofacsimile, comes under this heading. Here again the sensitivity of this type of transmission to changes in level is much greater than that of telephone service, say, about ten times as great. The requirement of limiting sudden level changes to something less than ¼ db necessitates special facilities for photofacsimile channels.

Continuous but small (fractions of a decibel) changes in level are caused by hunting of the dynamic regulators which are used in the newer types of telephone systems. These level changes can be detected in areas of what should be uniform middle tones in reception by photographic processes, and the effect has been appropriately called *uneasy background.* They are of no consequence in telephone transmission but in a circuit specially engineered for facsimile transmission the telephone company might want to apply in some line sections pilot-tone level-compensator equipment designed

Table 1. Approximate Maximum Lengths of Nonloaded, Nonrepeatered Exchange Area Cables with an Attenuation Loss of 12 Db at 1,000 Cps

Gage	Type	Capacitance, μf per mile	Max length, miles	2,400-cps loss increase over 1,000 cps, db
22	BSA and CSA	0.082	8.1	7.5
24	ASM and CSA	0.072	5.5	6.1
24	DSM	0.084	5.2	6.3
26	ST and AST	0.069	4.5	6.5
26	BST	0.079	4.2	6.4

to eliminate this effect. This would remove not only the uneasiness of the background but also frequency patterns (a special kind of recurring level changes discussed in the next section) coming from power or telephone circuits, and all other level changes of as much as ±5 db, when such a change occurs within the compensated section of line. The system[1] utilizes a pilot frequency just outside of and above the frequency band used by facsimile machines of a maximum element rate of about 2,400 and will therefore compensate for level changes during the transmission. While the system has a fast enough response to compensate for a 180-cps frequency pattern (the third harmonic of 60-cps power interference) this is not fast enough to compensate for short hits (defined below) but it will have a limiting effect on long hits.

Sudden and transient changes in received signal level are called *hits.* These may vary from a small portion of a scanning line to several scanning lines depending on the duration of the hit and the linear speed of recording.

43. Noise and Frequency Pattern Interference. Noise in a telephone-message communication channel becomes either random specks and streaks, or more or less regular frequency patterns, in the facsimile record copy. The specks or short streaks (random noise) may be caused by vacuum-tube noise, atmospheric static, ignition noises, cross modulation or cross talk, etc., and the frequency pattern by intermodulation of extraneous frequencies such as those of power circuits, or testing or signaling frequencies used in adjacent telephone circuits with frequencies in the facsimile transmission ranges.

In the case of random noise the rms level at the input to the facsimile receiver should be at least 35 to 40 db below the level of the maximum picture carrier when high-grade photofacsimile with photographically processed recordings is involved.

In black-and-white transmissions when copy appearance is not important but perfect legibility is demanded, the noise could be many decibels higher than this, depending, for instance, on the size of the lettering of the copy. For 12-point boldface characters the rms noise should be not less than 10 db below the signal.

[1] Jones, T. A., and W. A. Phelps, "A Level Compensator for Telephotograph Systems," AIEE Technical Paper No. 53-297, June, 1953.

Frequency pattern interference is important mainly in high-grade photofacsimile in which the copy may be so sensitive to this type of interference that it may be discerned even when it is 55 db or more below the maximum signal level.

44. Echo. The effect of echo on the appearance of a facsimile recording when the echo has a short delay is practically undistinguishable from the effects of delay distortion, although the causes are different. An echo may be caused by two approximately equal and sufficiently large impedance irregularities or one very large and one small one in the transmission path, separated from each other by a finite amount of delay, or it may be caused by having the main transmission path from the transmitter to receiver paralleled in part or entirely by a second path of a different amount of delay. In either case two (or more) sets of the same transmitted data are recorded with differences in time. Echo is a frequent and important cause of unsatisfactory

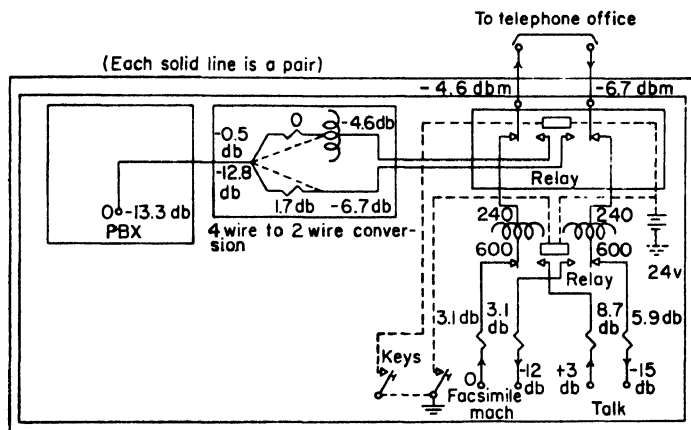



FIG. 36. Typical alternate facsimile full-period station arrangement.

copy in the case of high-grade phototransmission and may also, when severe, interfere with legibility of black-and-white transmissions.

45. Circuits Engineered for Facsimile Service. All common telephone-message circuits are "two-wire" circuits in at least some links connecting the two telephone sets. Many circuits are completely "two-wire" for the entire length. This means that the same transmission path is used for both directions of transmission. Such a circuit tends to have more noise, echo, and cross talk and to be much more vulnerable than a completely "four-wire" circuit where two separate paths, to a high degree independent of each other, are provided for each direction of transmission.

For more reliable, and generally more satisfactory facsimile transmission, the telephone circuit should be made effectively a four-wire circuit during the transmission of the facsimile signal. Other special conditioning is often required, such as delay and attenuation equalization, application of echo suppressors, level compensators, and maintenance within more stringent tolerances than those required for telephone service. For high-grade photofacsimile transmission such an upgrading of an ordinary telephone-message channel in most cases is a practical necessity and for black-and-white transmission it will frequently be found desirable.

46. Terminating Arrangements for Facsimile Service. There is a considerable variety of facsimile-service terminating arrangements, depending to some extent on local conditions and requirements at the place where the facsimile machines are installed.

In the case of a network with a number of different sending or receiving points, it is usually advantageous for both the facsimile-service user and the communications company responsible for maintaining the network for the terminating arrangements

at all points to be as nearly as practicable the same. An example of a terminating arrangement with no special features is shown in Fig. 36.

RADIO FACSIMILE

News photos and weather maps make up the greater part of facsimile radio traffic. There is some traffic in the form of financial reports and documents that cannot be readily transmitted by code or printing telegraphy. The problem of errors in these two systems is developing a trend toward facsimile for message traffic.

47. Radio-transmission Difficulties. Radio circuits making use of sky-wave transmission paths require the use of auxiliary equipment if the degradations are to be kept to a minimum. The basic problem of fading and noise is met by converting the a-m signal of the facsimile transmitter to a frequency shift signal of 800 cps total shift. This signal may be transmitted as an audio-frequency shift signal (AFS) or as a shift of the r-f carrier itself (RFS).

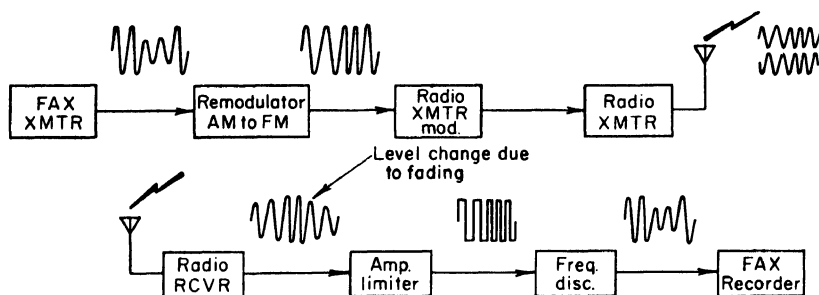



FIG. 37. Audio-frequency shift system using a voice-frequency radio circuit.

48. Audio-frequency Shift (AFS).[1] When a double-side-band a-m radio circuit or a narrow-band f-m circuit is available, AFS (formerly referred to as subcarrier frequency modulation) is most convenient. Both the transmitting and receiving stations are relatively easy to operate. The transmitting and receiving r-f oscillators do not require a high degree of stability. With respect to distance of transmission and freedom from multipath effects and noise, the AFS system is somewhat inferior to RFS transmission described later.

A remodulator is needed at the transmitter to convert the a-m facsimile signal into an AFS signal. A discriminator is required at the receiver to convert the AFS signal from the radio receiver into an a-m signal for the facsimile receiver.

The block diagram in Fig. 37 shows typical transmitting and receiving stations using AFS.

SSB a-m radio circuits are used with the same arrangement of the auxiliary equipment. The degradation caused by multipath phenomena is less than with the double-side-band a.m. or narrow-band f.m. If the carrier is entirely suppressed, the signal is the same as with the FS circuit.

The *remodulator* first demodulates the a-m carrier from the facsimile transmitter to obtain the base band. This signal may control the frequency of one oscillator in a beat-frequency oscillator (BFO) so the beat frequency at the output swings between 1,500 and 2,300 cps. In another system, the demodulated signal controls the pulse-repetition rate of a single-shot multivibrator. A third system which is simple and stable in operation employs an *RC* oscillator whose frequency is controlled by the demodulated signal.

This form of remodulator has an audio *RC* phase-shift oscillator which is more stable than

[1] Mathes, R. E., and J. N. Whitaker, Radio Facsimile by Sub-carrier Frequency Modulation, *RCA Rev.*, **4** (2), 131, October, 1939.

the BFO. The output frequency of the oscillator is varied by a tube in one of the legs of the frequency-determining *RC* network. Rectified signals from the facsimile transmitter are fed to this tube, which acts as a variable resistor to change the period of oscillation. The maximum input signal will produce the 1,500-cps output frequency (Fig. 38).

The *discriminator* units are composed of an amplifier, a limiter, and a frequency discriminator. The amplifier builds up the weakest workable signal to a level sufficient to reach the saturation point in the limiter. All signals above this level will be converted into square waves of constant amplitude. The output of the limiter passes through a frequency discriminator, the simplest type of which is a low-pass filter which passes signals in the 1,500-cps range with much less attenuation than the 2,300-cps signals. The contrast range of the output signal is generally adjustable to meet the requirements of the different types of facsimile receivers. The output signal of this discriminator still contains the a-f shift in addition to the amplitude modulation. Since most facsimile receivers are sensitive only to amplitude, it is usually not necessary to remove the frequency-shift component.

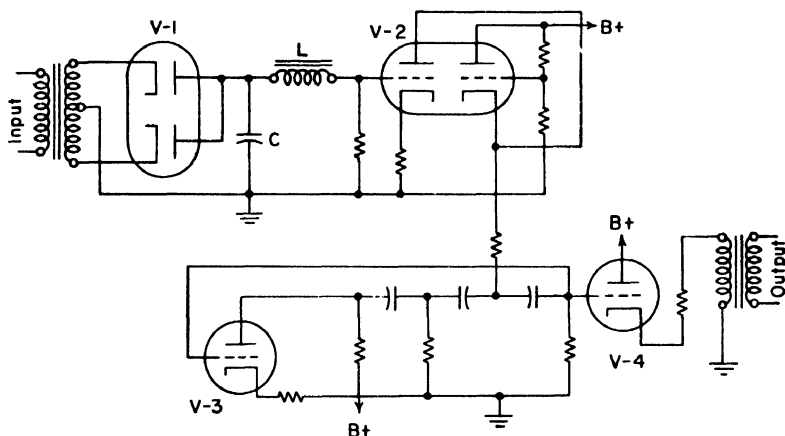



Fig. 38. Simplified schematic of remodulator.

Most other types of frequency discriminators demodulate the square-wave signal to the base band. This d-c signal modulates an a-f carrier to produce an amplitude-modulated carrier signal of fixed frequency. Examples of this type of frequency discriminator are: balanced phase-shift discriminator (as used in f-m broadcasting), frequency counter, single-shot multivibrator, etc. The output of the low-pass-filter frequency discriminator may be rectified to obtain the base band.

The amplitude of the input signal may vary over a range of 40 db or more without affecting the level of the output. Noise below the signal level will be lost in the square-wave envelope. Out-of-phase noise will put a reversal in the normal signal envelope and cause an unwanted marking on the recorded copy when the resultant signal drops below the clipping point.

49. Radio-frequency Shift (RFS). Facsimile signals can be transmitted over a regular c-w radio transmitter which is equipped for frequency-shift modulation. This system is generally used for permanent radio-facsimile circuits. The RFS system is superior in overcoming the effect of multipath signals and noise. Only a very narrow r-f band width is required.

Multipath radio signals, noise, and interference cause degradation of the received copy; so a radio frequency should be selected which minimizes these conditions.

It is customary to arrange a simple switching system so the radio transmitter can be quickly switched to either facsimile or teletype. The facsimile-transmission range is about the same as for teletype or c.w., unless some degradation can be tolerated; in that case, the workable transmission range for facsimile is greater.

The block diagram (Fig. 39) shows a transmitting and receiving station for an RFS radio circuit.

The auxiliary equipment required at the transmitting station is a demodulator to convert the facsimile signal into a varying d-c voltage for the frequency-shift exciter unit or keyer. The varying d-c voltage shifts the frequency of the radio transmitter in accordance with the amplitude of the d-c signal. The oscillators used at both the transmitting and receiving stations must be very stable.

The output of the radio receiver is an audio tone varying in frequency (AFS) in accordance with the frequency shift (RFS) of the radio carrier. The audio tone is developed as a beat frequency between the radio carrier and a stable r-f oscillator feeding the input of the radio receiver. On very stable radio receivers the BFO may be used instead of a stable r-f oscillator. This a-f shift tone is fed into a discriminator to convert it to an a-m signal before it is applied to the facsimile receiver. The discriminator eliminates the effects of signal fading and reduces the effects of interference and noise.

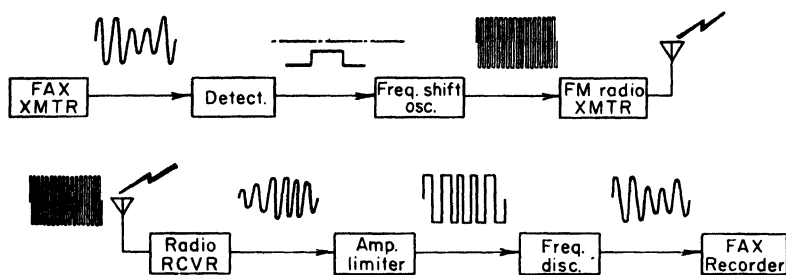



FIG. 39. General features of frequency-shift system.

FACSIMILE-MACHINE CHARACTERISTICS[1]

50. Compatibility. In order that facsimile equipment may be compatible, certain operating characteristics must be common to the transmitting and receiving equipments. Either the equipments must be synchronized to a common reference such as the power-line frequency, or synchronization must be maintained by using high-stability frequency standards at the transmitter and receiver. The degree of synchronization required will depend upon the amount of skew which is acceptable per copy. This requirement may vary from about 20 ppm for systems which have a relatively short total length of scanning line per copy to 0.2 ppm for systems which have a long total scanning line per copy.

If the *indexes of cooperation*[2] are not the same, the received copy will be stretched or compressed. Since it is only the aspect ratio of the received copy which is affected, the latitude permitted will depend upon the subject copy. Unless the difference in index is very great, typed or printed copy will be legible, but a photograph may be unacceptable. A difference in index of cooperation of as much as 10 per cent can often be tolerated.

For drum-type receivers the ratio of drum length to drum diameter (*drum factor*) must be at least as large as that of the transmitter to ensure that the end of the copy will not be cut off.

To obtain proper receiver phasing it must be designed to respond appropriately to the particular type of phasing signals sent out by the transmitter which is to operate it. Most facsimile systems use a series of pulses at the beginning of a transmission to modulate the carrier in the same manner as the picture signals. The

[1] By A. C. Mitchell, Times Facsimile Corp.

[2] (CCIR) The product of the scanning-drum diameter and the number of scanning lines per inch. IRE index = π times CCIR index.

pulses are usually shorter than 5 per cent of the total scanning-line length and are placed at the edge of the copy.

If control signals are used for automatic operation of the receiver it must be sensitive to the particular kinds of signals sent out by the transmitter. For use on circuits which may change the absolute frequency of the signal or the type of modulation, the control signals should modulate the carrier in the same manner as the picture signals.

For recorders which reproduce on an opaque record sheet, the same direction of scanning (normal or reverse) must be used at both ends. The receiver must also be capable of accepting the type of transmission (white transmission or black transmission) being used.

The receiver must be designed to operate on the type of modulation (amplitude or frequency shift) to be used and must have the proper contrast or frequency-shift limits. For good gray-scale reproduction the linearity of the signal amplitude or frequency with copy density must be known. In most systems designed for gray-scale reproduction the relation of signal voltage or frequency with copy density is approximately linear.

51. Compatibility Parameters. The operating characteristics of a facsimile system which determine the compatibility of the transmitters and receivers are listed below:

1. Stroke speeds (lines scanned or recorded per minute).
2. Direction of scan.
3. Starting operations if automatic starting and phasing are used; phasing operations if manual starting is used.
4. Index of cooperation. Wide tolerances are generally permissible for message or weather-map copy.

Table 2. Characteristics of Representative Facsimile Systems

	Copy size transmitted, in.	Drum diam, in.	Scanning lines per in.	Index of cooperation		Scans per min	
				CCIR	IRE		
A. Transmission of photographs	8 × 7	2.75	96	264	829	90 or 45	Armed Forces, R.C.A.C., International News photos, Mackay Radio, Press Wireless
	8 × 11.81	2.75	96	264	829	90	
	8 × 11.81	2.75	127	350	1,012	60	
	8 × 7	2.75	138	380	1,195	100	The Associated Press network
	8 × 11.75	3.80	100	380	1,195	100	
	8 × 10	2.75	96	264	829	90	International Newsphoto networks
	8 × 10	2.75	105	289	907	150	
	8.5 × 12	2.91	100	291	915	100	United Press networks
	8.5 × 12	2.91	150	435	1,370	120	
B. Transmission of meteorological maps and large documents	18 × 12	6.0	96	576	1,810	60	Armed Forces, Weather Bureau, and commercial airlines networks
	18 × 22	6.0	96	576	1,810	60 or 120	Canadian weather network
C. Transmission of business documents	8 × 13 or cont.	2.75	96	264	829	180 or 360	Bell system
	8 × 10	2.63	100	263	825	90, 180, or 360	Western Union
D. Special services.	In this category are systems designed to transmit special kinds and formats of subject copy such as telegrams, tickets, and X-ray films. Western Union has six different facsimile systems for public-service telegraph service with differing drum diameters, scanning lines per inch, and indexes of cooperation. It also has a system for transmitting tickets for space reservations, and for small documents such as brokers' orders. Times Facsimile supplies a facsimile system for transmitting 14- × 17-in. X-ray films at a drum speed of 180 rpm and an index of 264						

52. Facsimile-systems Standards. The International Telegraph Consultative Committee (CCIT) has set up the following specifications for facsimile machines as standards to be used in international communications:

For Photographs:
- Cylinder circumference ($D \times \pi$) .... 8.66 in.
- Min useful circumference .... 8.00 in.
- Max skew .... 0.15 in.
- Useful cylinder length .... 11.81 in.
- Max picture size .... 8.07 × 11.65 in.
- Drum factor .... 2.5 min
- Line advance (scanning density) .... 96 and 127 lines per in.
- Indexes of cooperation .... 264 and 350, respectively (830 and 1,100 IRE)
- Stroke speed (rpm for drums) .... 90 and 60, respectively, and multiples and submultiples

For meteorological charts and large documents:
- Index of cooperation (proposed) 576
- Drum speed (recommended) .... 60 rpm

Proposed for business documents:
- Index of cooperation 264
- Drum diam* .... 2.6 2.95 in.
- Min usable drum length of transmitters .... 11.8 in.
- Drum speeds of 90, 120, 150, 180, 240, and 360 rpm have been proposed

* U.S. State Department proposes standard of 2.75 in.

References

AIEE Technical Papers 52-103 and 56-108.
Bell System Tech. J., **4**, 187; **13**, 464; **15**, 549.
Department of the Army Technical Manual TM11-2258 (Department of the Air Force Technical Order to 31S3-2TXC1-1).
Hills, L., and T. J. Sullivan: "Facsimile," McGraw-Hill, 1949.
Instruction Book for Facsimile Recorder RD-92A/UX (Navships 91630).
Instruction Book for Facsimile Transceiver TT-41B/TXC-1B (Navships 91068A).
Instruction Book for Facsimile Transceiver TT-66A/TXC (Navships 91854).
Instruction Book for Recorder Set, Facsimile AN/FXH-3(XN-2) (Navships 91981).
Jenkins, C. Francis: "Vision by Radio," 1925.
Jones, Charles R.: "Facsimile," Murray Hill, 1949.
Korn and Glatzel: "Handbuch der Phototelegraphie und Telautographie," Otto Nemnich, Leipzig, 1911.
"Radio Facsimile," RCA Institutes Technical Press, 1938.
War Department Technical Manual TM11-375B.
Western Union Tech. Rev., **4**, 1, 10, 16; **6**, 92; **7**, 21; **8**, 90, 103, 144; **10**, 2, 174.

CHAPTER 24

CODE TRANSMISSION AND RECEPTION

By John B. Moore[1]

1. Basic Considerations. The transmission of information in coded form, rather than by normal voice communication or the written word, is accomplished by the use of so-called "signals" of one form or another. This may be the beating of drums, the use of flags, smoke signals, lanterns, flashing lights, or suitable variation or interruption of an electric current or of an electromagnetic wave. The nature of the information to be transmitted and the capabilities or limitations of the method and facilities available for accomplishing such transmission and reception determine the type of code and its detailed construction.

Suitable coding not only puts information into a form that can be transmitted by available means but it also can be employed in various ways to abbreviate a message or to provide secrecy.

Telecommunication over radio circuits generally must employ some form of coding. Voice communication may be considered as being basically a case of direct reproduction that does not involve any coding of the original intelligence (sampling and pulse coding methods are not considered here). Other forms of intelligence, however, require a coding scheme that will make possible the transmission of the desired information over the type of physical facilities, equipment, and circuit that are to be used.

CODES

Most codes, employed for signaling over wire or radio circuits, are basically of the so-called binary or two-condition type; these two conditions being designated by such terms as *mark-space, on-off, positive-negative, high-low, upper-lower*. Three-condition codes are used on submarine cables, while certain other systems utilize what may be considered *four-condition signaling*. In any such code, a code combination consists of alternate short or long periods of time during each of which only one of the two or more possible conditions is established.

A different type of coding is obtained by the simultaneous transmission, through separate electrical paths over the circuit, of two or more discrete frequencies or tones, each code character consisting of a different combination of one, two, or more of the possible number of frequencies (see Chap. 27).

2. Specific Applications. A full understanding of the fundamental role of coding, in the present-day telecommunications art, requires at least a listing of the various types of service that are based on suitable coding of the electrical signal.

Telegraphy. Codes, such as those shown in Figs. 1 and 2, assign a separate code combination to each letter, character, special function, or arbitrary instruction to be transmitted. Such codes, for radio telegraphy (RT), generally are of the so-called binary or two-condition type.

Switching. Remote or selective control of functions and of communications circuits or routings generally is effected by use of telegraph-type binary codes or of simultaneous multifrequency codes.[5*]

Telemetering. Where measurements of physical quantities are to be automatically transmitted to a distant point, it generally is necessary to convert the relative or

[1] Engineering Department, RCA Communications, Inc.

* Superior numerals refer to References at end of chapter.

Character	Code
A	.-
B	-...
C	-.-.
D	-..
E	.
F	..-.
G	--.
H	....
I	..
J	.---
K	-.-
L	.-..
M	--
N	-.
O	---
P	.--.
Q	--.-
R	.-.
S	...
T	-
U	..-
V	...-
W	.--
X	-..-
Y	-.--
Z	--..
Ä (German)	.-.-
Á or À (Spanish-Scandinavian)	.--.-
CH (German-Spanish)	----
É (French)	..-..
Ñ (Spanish)	--.--
Ö (German)	---.
Ü (German)	..--
1	.----
2	..---
3	...--
4	....-
5	.....
6	-....
7	--...
8	---..
9	----.
0	-----

Character		Code
Period	.	.-.-.-
Comma	,	--..--
Colon	:	---...
Question mark, or request for repetition of a transmission not understood	?	..--..
Apostrophe	'	.----.
Dash or hyphen	–	-....-
Fraction bar	/	-..-.
Parenthesis (before and after words)	( )	-.--.-
Quotation marks (before and after words)	"	.-..-.
Equal sign	=	-...-
Understood		...-.
Error		....
Cross or end-of-telegram or end-of-transmission signal		.-.-.
Invitation to transmit		-.-
Wait		.-...
End of work		...-.-
Starting signal (beginning every transmission)		-.-.-

FIG. 1. International Morse code teleprinter codes and typical character assignments.

numerical values of such measurements into a suitably coded form for electrical transmission over a wire or radio circuit, the appropriate type of code being dependent upon information requirements and upon the type of communication channel.

Pulse Modulation. This is basically a binary or two-condition type of coding, in which the original intelligence is sampled at suitably frequent intervals and then is converted into pulses which are so modulated, positioned in time, or coded, as to make possible their reconversion into the original form of the intelligence being transmitted.

Facsimile. The electrical transmission of black-and-white copy and of photographs containing all shades of gray ranging from black to white involves scanning and conversion processes that may be considered a form of coding.

Television. The control of scanning, synchronizing, color selection, etc., necessarily is based on the use of a standard coded signal, the resultant television signal itself being a coded form of the total intelligence being transmitted.

TELEGRAPH CODES

The following treatment is limited essentially to those codes authorized by the international Telegraph Regulations, Paris, 1949,[1] for use over international wire-line and radio circuits. These are so-called binary or two-condition codes, in which the time duration of any marking or spacing interval is a multiple of the shortest or unit interval or unit element (sometimes referred to as a Baudel). Descriptions of other codes, which are of specialized types or which enjoy less widespread use today, can be found in the literature.[2,3]

3. Morse Code. This type of code, one example of which is shown in Fig. 1, consists of dots, dashes, and spaces. The dot is the shortest, or unit, element. A space between dots and/or dashes is equal in length to a dot. A dash is equal in length to three dots. A space between letters or characters is equal in length to three dots, and a space between words to seven dots. (When transmission is by perforated tape, the space between two letters is equal to one "center hole" and the space between two words is equal to three "center holes.")

NOTE: The term "dot," as used here, denotes a mark (M) element of unit length. It should not be confused with the term "dot-cycle," which denotes a mark unit element followed by a space (S) unit element.

The international Morse code, specified in the Telegraph Regulations,[1] is shown in full in Fig. 1. Accented letters peculiar to certain languages other than English, shown in a separate block on the code chart, are used only between offices, stations, or ships of such countries as permit their use. Nationals of countries using non-English syllabaries or alphabets, such as the Japanese and Russian, must spell out the sounds of their own words in one of the languages using the English alphabet. This sometimes requires the use of arbitrary equivalents.

The American Morse code, used chiefly on domestic telegraph services within the United States, differs from the international Morse code in the code combinations assigned to certain letters and to the numerals, and in the use of one lengthened space within the internal structure of a few of these code combinations.

Japan, Russia, and other countries that have their own national syllabaries and alphabets—differing radically from the English alphabet—employ special forms of the Morse code[4] for working within their own countries and territories and to their ships.

4. Morse Cable Code. This differs from the international Morse code, of Fig. 1, in employing equal-length pulses of three polarities—positive, negative, and neutral—to represent a dot, a dash, or a space between letters, respectively. Therefore, it is considerably faster than the dot-dash-space form of the two-condition Morse code. However, the limitations of radio reception, and of most RT transmitters prior to the advent of modern methods of frequency-shift keying, prevented the use of such a three-condition code for radio telegraphy.

To make available the advantages of this equal-element type of cable code, in RT operations, the British Higgitt system employs a so-called two-condition cable code

in which a Morse dot is represented by a M-to-S transition, a Morse dash by a S-to-M transition, and a space between letters by the absence of a transition in the radio signal. The code can be used only in a synchronous system, such as a time-division multiplex (TDM) system, especially designed to function in accordance with precise

5-Unit Baudot Code					LETTERS	FIGURES							ARQ 7-Unit Moore Code						
						International	W. U.	Military	TWX	Weather									
1	2	3	4	5									1	2	3	4	5	6	7
O	O	–	–	–	A	–	–	–	–	↑			–	–	O	O	–	O	–
O	–	–	O	O	B	?	?	?	5/8	⊕			–	–	O	O	–	–	O
–	O	O	O	–	C	:	:	:	1/8	O			O	–	–	O	O	–	–
O	–	–	O	–	D	Idf.	$	$	$	↗			–	–	O	O	O	–	–
O	–	–	–	–	E	3	3	3	3	3			–	O	O	O	–	–	–
O	–	O	O	–	F	Opt.		!	1/4	→			–	–	O	–	–	O	O
–	O	–	O	O	G	Opt.	&	&	&	↘			O	O	–	–	–	–	O
–	–	O	–	O	H	Opt.	#	Stop	Stop	↓			O	–	O	–	–	O	–
–	O	O	–	–	I	8	8	8	8	8			O	O	O	–	–	–	–
O	O	–	O	–	J	Bell	Bell	'	,	↙			–	O	–	–	–	O	O
O	O	O	O	–	K	(	(	(	1/2	←			–	–	–	O	–	O	O
–	O	–	–	O	L	)	)	)	3/4	↖			O	O	–	–	–	O	–
–	–	O	O	O	M	.	.	.	.	.			O	–	O	–	–	–	O
–	–	O	O	–	N	,	,	,	7/8	⦶			O	–	O	–	O	–	–
–	–	–	O	O	O	9	9	9	9	9			O	–	–	–	O	O	–
–	O	O	–	O	P	0	0	Ø	0	Ø			O	–	–	O	–	O	–
O	O	O	–	O	Q	1	1	1	1	1			–	–	–	O	O	–	O
–	O	–	O	–	R	4	4	4	4	4			O	O	–	–	O	–	–
O	–	O	–	–	S	'	'	Bell	Bell	Bell			–	O	–	O	–	O	–
–	–	–	–	O	T	5	5	5	5	5			O	–	–	–	O	–	O
O	O	O	–	–	U	7	7	7	7	7			–	O	O	–	–	O	–
–	O	O	O	O	V	=	;	;	3/8	Φ			O	–	–	O	–	–	O
O	O	–	–	O	W	2	2	2	2	2			–	O	–	–	O	–	O
O	–	O	O	O	X	/	/	/	/	/			–	–	O	–	O	O	–
O	–	O	–	O	Y	6	6	6	6	6			–	–	O	–	O	–	O
O	–	–	–	O	Z	+	"	"	"	+			–	O	O	–	–	–	O
–	–	–	O	–	Carriage Return								O	–	–	–	–	O	O
–	O	–	–	–	Line Feed								O	–	O	O	–	–	–
O	O	O	O	O	Letters								–	–	–	O	O	O	–
O	O	–	O	O	Figures								–	O	–	–	O	O	–
–	–	O	–	–	Word Space								O	O	–	O	–	–	–
–	–	–	–	–	Blank								–	–	–	–	O	O	O
					———							RQ Signal	–	O	O	–	O	–	–
					———							Idle Alpha	–	O	–	O	–	–	O
					———							Idle Beta	–	O	–	O	O	–	–

Opt. = Optional

Idf. = Identification / Answer Back / Who Are You

FIG. 2. Printer-key arrangements and codes.

timing and other requirements imposed by the code and its proper interpretation by the receiving terminal equipment.

5. Teleprinter Codes. These are distinguished by the feature that, unlike the Morse code, each combination in a given code is of the same length as measured in unit elements or in milliseconds of time duration. Typical examples of the teleprinter codes employed over radio circuits are the five-unit Baudot code and the

"protected" seven-unit constant-ratio code[6] shown in Fig. 2. Both are shown in their basic form as used in synchronous systems where the transmitting and the distant receiving machines (or their electronic equivalents) run continuously and are automatically maintained in synchronism in speed and phase. For nonsynchronous or so-called start/stop operation, where the distributor mechanism of the receiving machine must be started and stopped for each code combination transmitted, each basic combination shown in the chart of Fig. 2 must be preceded by a start (space) signal of unit-interval duration and must be followed by a stop (mark) signal which, in different designs of system, may have a duration of 1.0 or 1.42 or 1.5 unit elements. Thus, a 5-unit start/stop code becomes actually a 7.0- or 7.42- or 7.5-unit code. Any other basic teleprinter code, such as the 6-unit or 7-unit or 8-unit, may similarly have added to it the start and stop intervals required to adapt it for start/stop operation.

6. Baudot Nonprotected. This equal-length teleprinter code, in its widely used five-unit form, is given in Fig. 2. (Start and stop intervals for start/stop operation are not shown.) All combinations are of the same length or time duration, each consisting of five unit elements. By assigning to each unit element either M (0) or S (−) status, there is obtainable a total of 32 combinations. Typical assignments to letters, machine functions, and other characters are shown in Fig. 2, where the letters assignments designate the lower-case position or shift of the receiving printer, and figures assignments designate the upper-case position or shift. A six-unit code, of this type, provides a total of 64 combinations, examples of its application being in Teletypesetter equipment and in the Japanese domestic teleprinter network.

This is called a "nonprotected" code because any mutilation of the signal which changes a mark to a space, or vice versa, results in a different code combination and thus causes an error in the received message.

7. Constant-ratio Protected.[6] This equal-length teleprinter code, shown in its presently used seven-unit form in Fig. 2, is so constructed that each combination consists of 3 M and 4 S unit elements, giving a constant M/S ratio of 3:4. The total number of such combinations is 35. Three combinations, as indicated in the chart of Fig. 2, are assigned to special functions associated with a particular type of system in which this code is used for RT services. An eight-unit code of this type provides 70 protected combinations.

This code provides a high degree of protection against errors caused by mutilated signals, since any change from the correct 3:4 ratio of M to S elements results in an invalid combination that can be detected and registered as such at the receiving office. In the case of weak signals severely affected by noise and fading, it is possible that one unit element may be changed from M to S and another from S to M within one code combination. Such a compensating double mutilation results in a so-called "transposition," the resultant combination having a valid 3:4 ratio, but not being the combination that was transmitted. Fortunately, the probability of occurrence of such transpositions is small. On accuracy of transmission, see Chap. 27.

8. Error-correcting Codes.[7,8] Usually these are two-condition codes in which each combination is so constructed that the occurrence of a specified type of single or multiple mutilation—one or more unit elements changed—will not only be detected but can be interpreted by the receiving decoder or converter in such a manner that there is a high probability that the signal mutilation will not result in an error in the recorded information. This error-correcting feature is obtained by adding, to some or all of the basic code combinations, one or more so-called "check digits" or unit elements as required (see Chap. 27).

This type of code finds application in signaling or data-handling systems where its use may be preferable to manual or automatic request for a repetition of the mutilated characters. It is not used in commercial RT services, where the constant-ratio type of protected code is finding increasing application.

9. Code Conversion.[3,9,10] When it is necessary to transfer message traffic or other information from a system employing one code to a system employing a different code or type of code, the transfer may be effected either by manual transcription or

by automatic code conversion. The trend is to the use of automatic converters employing mechanical selection, electromechanical relays, electronic devices and circuits, or matrixes of various types. A few typical cases, of commercial importance, are mentioned below.

10. Morse and Five-unit.[9] This is a particularly difficult type of conversion because, as will be seen from the code charts of Figs. 1 and 2, the five-unit teleprinter code makes use of letters (lower case or shift) and figures (upper case or shift) assignments, whereas the Morse code provides only one assignment for each code combination.

11. Five-unit Variants. As shown in Fig. 2, assignments to letters and to machine functions are well standardized but the figures assignments vary considerably among keyboards used in various services. An outstanding case of considerable commercial importance is that of the International assignments employed in the European Telex service and overseas TEX (RCA) service vs. the assignments employed by the TWX service in the United States. Since a number of the figures appearing in one do not appear at all in the other, it is impossible to provide a complete character-for-character conversion. Such a case requires, in addition to direct conversions, mutual agreement on a number of suitable equivalences and arbitrary operating procedures whereby the transmitted intelligence can be correctly conveyed in the received copy.

12. Five-unit and Seven-unit.[3] This is a commercially important example of conversion between a protected and a nonprotected type of teleprinter code, the seven-unit protected code being used over many international radio circuits, and the five-unit nonprotected code in commercially available teleprinter equipment and over wire-line or other forwarding circuits. As will be seen from Fig. 2, conversion between the two codes requires merely a conversion between five-unit and seven-unit code combinations assigned to the same letter or function. The presently used assignment of seven-unit combinations was arranged specifically to facilitate conversion between these seven- and five-unit codes.

13. Business Codes. Business concerns that have a large volume of telegraphic communication commonly use standard "commercial codes" consisting of groups of letters arranged alphabetically, each group standing for a word, phrase, sentence, or other item of information. Typical examples are the ABC Code (6th edition), Bentley's, and Peterson International Code. There also are more specialized codes for certain industries or branches of commerce; and some concerns employ private codes to provide secrecy in addition to abbreviation. Morse press-wire operators used a standard list, known as the Phillips code, of extreme abbreviations of common English words.

KEYBOARDS

From Figs. 1 and 2, it will be apparent that keyboards of Morse-code perforators, and of teleprinters used in various services, must differ appreciably from one another and from the keyboard of a standard office typewriter. Keyboards for the international Morse code and for international teleprinter services are shown in Fig. 3. While the basic arrangement of the letter keys follows that of a standard office typewriter, the key assignments for numerals and other figures vary widely. This applies not only within the United States but also to machines used in a number of other countries.

Machines employing the six-unit Baudot code instead of the more usual five-unit have enlarged keyboards to accommodate the greater number of characters made possible by the available 64 code combinations. An example is Teletypesetter equipment. This keyboard is shown in Fig. 3 with a number of special characters and functions deleted to facilitate comparison with the other keyboards. The lower-case or unshift position of letter keys provides small letters, and the upper-case or shift position provides capital letters since, unlike telegraph practice, both are required in typesetting.

MODES OF OPERATION[11,13]

14. Morse. In RT services of the world, earlier art dictated the use of manual Morse operation in which the transmitting operator sends with a hand key and the

MORSE PERFORATOR
Creed Type 9W

5-UNIT TELEPRINTER
International No. 2

6-UNIT TELETYPESETTER

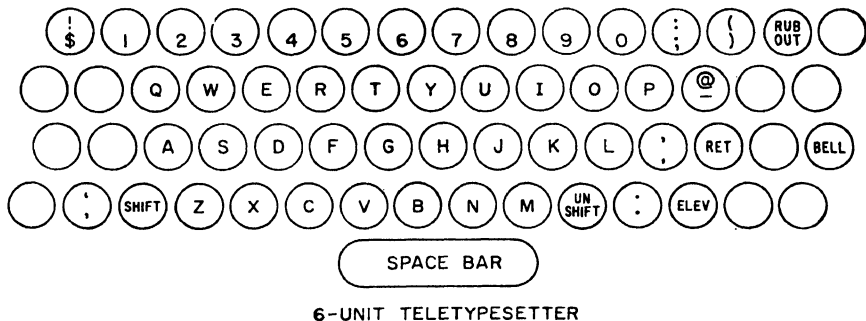
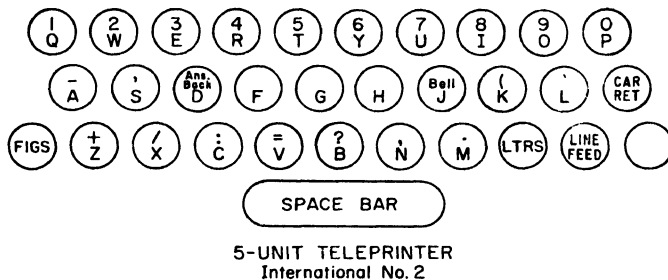
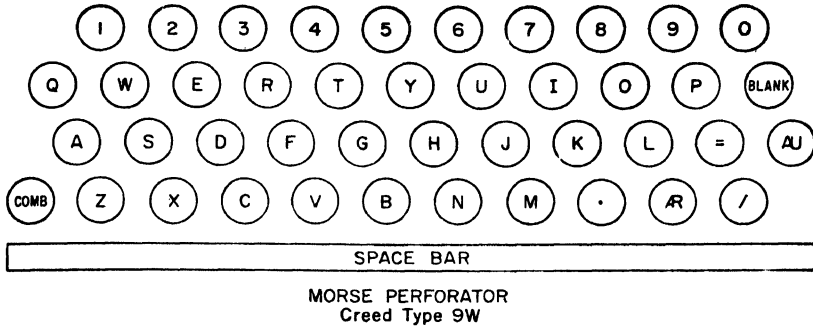



FIG. 3. Perforator and teleprinter keyboards. The six-unit Teletypesetter keyboard has 60 keys plus space bar. Assignments over numerals differ depending on service. Keys not marked provide special typesetting functions.

receiving operator copies by ear. This mode of operation still is used where necessitated by considerations of signal-to-noise ratio, practical operations, and economics. Point-to-point overseas commercial services, however, use mechanized Morse operation in which transmission is from perforated tape prepared by the transmitting operator, and reception is by reading and transcription from an up-and-down dot-dash ink record on a moving paper tape from a so-called ink recorder or undulator. Operating speeds may run as high as 300 or more words per minute (wpm).

15. Morse Direct Printer. To eliminate the necessity for manual transcription by a receiving operator, use has been made of a printing telegraph machine that operates directly from Morse code signals. A notable example is the Higgitt DCCC (double-current or two-condition cable code) system, still used by the British on their overseas RT network.

16. Teleprinter.[12] The use of teleprinter equipment employing equal-length codes of the Baudot and the constant-ratio types for commercial RT services has developed rapidly since about 1932. Among the reasons for this world-wide trend are availability of equipment, the need for mechanized operation and tape relay,[12] the difficulty of training Morse operators, protection afforded by the constant-ratio type of code,[6] and compatibility of structure of the nonprotected and protected codes.

17. Synchronous Operation. This does not require the addition of start and stop intervals preceding and following each combination of the basic codes shown in Fig. 2, but it does require relatively complex and expensive equipment at the terminal offices, and therefore is utilized chiefly on those overseas services on which the volume of traffic warrants the cost of the necessary equipment or where the particular service requires the use of protected-code operation such as the Automatic-RQ method discussed in Sec. 21. Synchronous terminal equipment generally is designed to provide two-channel or four-channel time-division multiplex (TDM) operation at approximately 60 wpm per channel. While special types of receiving printers have been used in such synchronous operations, present practice is to terminate each channel in standard start/stop terminal equipment, such as receiving printers or reperforators, or to extend one or more channels over wire-line or other start/stop circuits.

18. Start/Stop Operation (see Chap. 27). Where traffic volume or other service requirements neither dictate nor warrant the use of synchronous equipment, it is customary to employ the standard start/stop teleprinter machines and facilities widely used on wire-line services. This provides a single channel, at nominal speeds ranging from 60 to 100 wpm.

19. Nonprotected Operation. Standard types of five- and six-unit start/stop teleprinter equipment may be used on a single-channel basis over most radio facilities or may be used directly on the individual channels of time-division or frequency-division multiplex systems over important radio circuits. In either case, signal mutilations will cause errors in the received copy because of the nonprotected structure of the Baudot five- and six-unit codes.

20. Protected Operation. Commercial telegraph and other services often are called upon to provide a maximum of protection against errors due to signal mutilations. This is of particular importance over short-wave circuits, where signal-to-noise ratio and fading conditions vary greatly, and in any type of direct customer-to-customer service such as Overseas Teleprinter Exchange Service (TEX), leased channels, etc. Such protection appears to be provided most satisfactorily by the use of a constant-ratio type of protected code such as the 3 M–4 S form of seven-unit code described before. Such channels now are operated in so-called *diplex* pairs, on a two-channel synchronous basis, as required by the usual designs of Automatic-RQ (ARQ) equipment.[10]

21. Automatic-RQ.[10] This is a form of teleprinter operation employing a constant-ratio code in which detection of a mutilated character automatically prevents the receiving printer from functioning, initiates a request for repetition (RQ), stops transmission of traffic until the repetition is automatically provided from storage at the transmitting office and is correctly received, and then resumes normal transmission and reception. Its use requires that the two-way radio circuit be operable and functioning in both directions. Present-day designs of ARQ equipment operate such protected channels, in diplex pairs, over a two-channel or four-channel TDM system.

22. Multiplex. Whereas mechanized Morse operation made it possible to operate a good radio channel at speeds of 300 or more wpm, the use of teleprinter equipment limits each such channel to a speed of some 60 to 100 wpm. Efficient utilization of available frequency assignments and of radio transmitters, etc., makes it desirable and often necessary to employ so-called multiplex methods that permit the simultaneous operation of two or more teleprinter channels over one transmitter and frequency assignment. The two basic methods are time division and frequency division, combinations of the two systems sometimes being employed.

23. Time-division Multiplex[14,15,16] (see Chap. 27). Simultaneous transmission of two or more channels (messages) over a single carrier is effected by dividing the time duration of each single-channel unit interval into generally two, three, or four equal parts—for two-, three-, or four-channel operation—and utilizing these half- or third- or quarter-length intervals as unit elements of the multichannel, or so-called multiplex, aggregate signal (for details regarding modes of channeling, see Sec. 70). Each such aggregate-signal unit element carries the mark or the space intelligence of a full single-channel unit element of one of the individual channel signals. The multichannel aggregate keying is applied to a tone-frequency carrier, or to an RT transmitter, suitable for such high-speed keying as 120 or 240 wpm. At the distant receiving office, synchronous equipment breaks down the aggregate keying and delivers regenerated, single-channel signals to the separate receiving teleprinters for the individual channels. The number of channels is limited, in practice, by the maximum keying speed that can be handled satisfactorily over the facilities and circuits involved; multipath propagation on long-distance short-wave circuits generally sets a practical upper limit of four such TDM channels of nominally 60 wpm each.

24. Frequency-division Multiplex. The best-known form of FDM is the multitone voice-frequency carrier telegraph system employed over wire-line telephone networks and single-side-band (SSB) radio-telephone circuits, channel mid-band frequencies being spaced 120, 150, or 170 cps in different commercial systems. Each discrete tone carries the keying of a separate, start/stop teleprinter, all channels operate simultaneously and are separated at the terminal offices by electrical filters. Such a system provides flexibility and simplicity of direct interconnection between wire-line and radio services that employ start/stop teleprinter equipment of compatible design. Since the system does not involve time division, the maximum keying speed that must be handled over the radio circuit is that of any one tone channel. Where protected teleprinter operation is desired, the time-division diplex aggregate keying of a pair of ARQ channels may be applied to any tone of an FDM system, assuming there is adequate channel band width. In practice, the number of tone channels is limited by the reduction in received signal-to-noise ratio resulting from the necessity of reducing the power per channel on any given radio transmitter of fixed peak power rating as the number of channels is increased.

25. Twinplex.[17,18] This is a special form of two-channel multiplex in which channel distinction is achieved not by time division but by frequency separation in a special application of frequency-shift keying (FSK); four discrete frequencies are used to represent the four possible combinations of simultaneous M and S on the two channels. Keying of the two channels need not be, but preferably is, synchronized. More than two channels may be provided over one radio transmitter by extension of the basic method or by combining it with others.

KEYING SPEEDS

26. Definitions

Dot. The shortest (mark) element in codes such as the Morse code; the unit element in terms of which the length or time duration of dots, dashes, and spaces is measured.

Dot-cycle. A mark unit interval followed by a space unit interval, a dot-cycle per second being the unit in which keying frequency is expressed.

Unit Element or Unit Interval. In codes such as the Morse telegraph and the Baudot teleprinter code, the shortest or unit element or time interval used in forming code characters, each code character of the basic Baudot code consisting of five such unit elements or unit intervals.

Baudel. A term proposed, though not yet internationally accepted, to designate one unit element or unit interval in a telegraph or teleprinter code.

Baud. A unit of telegraphic signaling speed. The speed in Bauds is the number of unit elements, or unit intervals, per second, being numerically equal to twice the fundamental keying frequency expressed in dot-cycles per second. (A speed of 50 Bauds is equivalent to 25 dot-cycles per second.)

Line	Code	Sync or S/S	No. of channels	Total unit intervals per character	Unit intervals in stop element	Millisec per signal element	Millisec per stop element	Baud speed	Equivalent keying frequency, cps	Opm	Wpm 5 letters and 1 space	Remarks
1	5U	S/S	1	7.5	1.5	20.0	30.0	50.0	25.0	400	66 2/3	New CCIT standard
2	5U	S/S	1	7.42	1.42	33.7	47.9	29.68	14.84	240	40	
3	5U	S/S	1	7.42	1.42	22.0	31.2	45.5	22.75	368	61 1/3	TWX
4	5U	S/S	1	7.42	1.42	17.57	25	56.88	28.44	460	76 2/3	
5	5U	S/S	1	7.42	1.42	13.5	19.2	74.2	37.1	600	100	
6	5U	S/S	1	7.42	1.42	11.0	31.2	91.0	45.5	736	122 2/3	Teletype diplex
7	5U	S/S	1	7.42	1.42	20.0	28.4	50.0	25.0	404	67 1/3	European equipment
8	5U	S/S	1	7.0	1.0	23.3	23.3	42.93	21.46	368	61 1/3	
9	5U	S/S	1	7.0	1.0	22	22	45.5	22.75	390	65	WU domestic tape and page printer system
10	5U	S/S	1	7.0	1.0	20.0	20.0	50.0	25.0	428 4/7	71 3/7	Telex, Europe
11	6U	S/S	1	8.5	1.5	11.8	17.7	85	42.5	600	100	Stock ticker
12	6U	S/S	1	8.0	1.0	20.4	20.4	49.1	24.5	368	61 1/3	Teletypesetter
13	6U	Sync	2	6.0	....	13.33	.....	75	37.5	750	125	Teletype multiplex
14	6U	Sync	3	6.0	....	8.89	.....	112.5	56.25	1,125	187 1/2	Teletype multiplex
15	6U	Sync	4	6.0	....	6.67	.....	150	75	1,500	250	Teletype multiplex
16	7U	Sync	1	7.0	....	23.3	.....	42 6/7	21 3/7	367 17/49	61 11/49	RCA MUX
17	7U	Sync	2	7.0	....	11.7	.....	85 5/7	42 6/7	734 34/49	122 23/49	RCA MUX, TOR as used by RCA
18	7U	Sync	2	7.0	....	10.0	.....	100	50	857 1/7	142 6/7	TOR (European standard)
19	7U	Sync	4	7.0	....	5.8	.....	171 3/7	85 5/7	1,469 19/49	244 44/49	RCA MUX
20	DCCC	Sync	2	18 c.t.h. per word	....	.....	.....	75	37.5	....	125	Higgitt

FIG. 4. Teleprinter speeds and timing.

Operations per Minute (Opm). In teleprinter operation, the number of letters, characters, or machine operations per minute. This is the number of teleprinter code combinations transmitted per minute.

Words per Minute (Wpm). In computing telegraph transmission speeds, a "word" is considered as consisting of five letters and one space, or a total of six code characters.

27. Relative Speeds.[2,3] The relationship between Baud and dot-cycle per second keying speeds, and wpm traffic speeds for various codes, may be tabulated as follows:

For 100 Baud Speed or 50 Dot-cycles/Sec

Code	Wpm
5-unit synchronous	200
2-condition cable code	143
7-unit synchronous	143
5-unit (7.5) start/stop	133
Morse	111*

* The value for Morse operation is empirical because of the differing lengths of Morse code characters. It is based on the average context of commercial telegrams. Another conversion factor often used is that 40 dot-cycles per sec is equivalent to 100 wpm Morse.

28. Speeds Attainable. These are determined by the mode of operation, signal-to-noise ratio, fidelity of M/S ratio, and traffic requirements.

Morse. Manual operation ranges roughly from 15 to 35 wpm, depending on signal conditions and on operator proficiency. In mechanized operation speeds normally range from 30 up to 300 wpm depending on signal conditions, equipment capabilities, and traffic requirements.

Morse Direct Printer. In the British Higgitt system, the operating speed of the synchronous equipment normally is fixed at 125 wpm total for the two channels (see Fig. 4, last line).

Teleprinter and Multiplex. The tabulation of Fig. 4 gives, for various types of teleprinter equipment and services, common operating speeds expressed in Bauds, dot-cycles per second equivalent keying frequency (cps), operations per minute (opm), and words per minute (wpm).

KEYING ANALYSIS AND BAND WIDTHS

29. On-Off Keying.[19,20,21,26] All telegraph keying basically is on-off. In most, if not all, systems the channel keying originates as two-condition on-off keying and although it may undergo transformations it must end up as a reasonably accurate reproduction of the original keying in the two-condition mode. Therefore the two-condition rectangular wave form of voltage or current is the basis for analyses of keying side bands, occupied and required band widths, etc. (Special cases, such as the "shaped" keying of on-off radio transmitters with class C power amplifiers, require analysis of trapezoidal or similar keying wave shapes.[20]) Rigorous mathematical treatment would require a transient analysis of individual unit elements or of longer intervals. For practical purposes, however, it is satisfactory to consider that a continuous and unbroken succession or "train" of rectangular dot-cycles represents the maximum keying-frequency components and band-width requirements.

30. Fourier Analysis. In a dot-cycle, the value of the current or voltage rises instantly to a steady on or M value, at which it remains for one half cycle, and then instantly drops to a steady off or S value and remains there for the following half cycle. A wave train of such rectangular dot-cycles may be analyzed as consisting of the fundamental keying (dot-cycle) frequency plus all odd harmonics thereof. The equation of the voltage wave is

$$e = \frac{4E}{\pi}\left(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots\right)$$

which holds for values of x between $-\pi$ and $+\pi$. The first term, $\sin x$, represents the fundamental keying or dot-cycle frequency and, alone, provides merely a sinusoidal wave shape. The third, fifth, and succeeding terms, or harmonics of the

fundamental keying frequency, progressively steepen the rise and fall of the wave shape and flatten its top; the contribution of these terms becomes successively less significant as the order of the term increases. In practice, the fifth harmonic generally is the highest that is justified and that need be retained for a satisfactorily close approximation to the rectangular wave shape.

NOTE: Actual telegraphic keying contains M and S intervals ranging randomly from one to seven unit elements in length. The distribution of energy over the frequency spectrum therefore is not confined to the discrete odd-harmonic frequencies of a dot-cycle train but is more random within approximately the same frequency limits.

31. Band Width Occupied. Application of such a dot-cycle train of keying to a voice-frequency (v-f) tone carrier or to an r-f carrier produces the usual double-side-band signal of conventional amplitude modulation, the side bands extending out on each side of the carrier to a distance or frequency separation determined by the fundamental keying frequency and the number of its harmonics present in the keying wave shape or r-f envelope. Passage of the third harmonic, for example, results in an occupied band width of six times the fundamental (dot-cycle) keying frequency. The economics of wire-line operation, the growing congestion and demand for channels in the radio spectrum, and the necessity for reduction of interference and cross talk all dictate that the occupied band width be held to a minimum for the particular service. This requires that the rise and fall of the M-S and S-M transitions be made less abrupt and less steep or, viewed differently, that the higher-order harmonics of the keying frequency be suppressed. This is accomplished by the use of suitable low-pass filters or band-pass filters or of simpler *smoothing circuits*. In the case of most transmitters, for on-off keying, the practical limit of such reduction is a compromise between what is technically desirable and what is economically feasible.[21,26]

32. Signal Shaping. Shaping of the transmitted radio signal, as recommended by the CCIR (Geneva, 1951),[21] is specified as follows:

For single channel, a-m, c-w telegraphy (class A1) using Morse code, when large fluctuations of the field strength are present:

2.1.1 *Band width necessarily occupied.* The band width necessarily occupied is equal to five times the keying speed in Bauds (later referred to as B) with an attenuation of the components at the edges of the band equal to at least 3 db, in comparison with the level of the same components of the spectrum representing a series of equal rectangular dots and spaces at the same keying speed. This relative level of −3 db corresponds to an absolute level of 27 db below the level of a continuous mark.

2.1.2 *Spectrum of the out-of-band radiation.* Outside the band width defined above, the envelope of the spectrum should lie below a curve starting at the points (±5B/2, −27 db) defined above, and presenting a slope of 30 db per octave and extending over at least one octave, that is, out to the points (±5B, −57 db). From these points onward, the level of all components emitted should be below −57 db.

2.1.3 *Build-up time of the signal.* The spectrum limited precisely as described above corresponds to a build-up time of the signal equal to about 15 per cent of the initial duration of the telegraph dot. This build-up time is likely to vary slightly around that figure, depending upon the exact nature of the filters employed for the shaping of the signals.

33. Band Width Required. This is determined by the fundamental keying frequency—in dot-cycles per second—and the steepness of M-S and S-M transition required by the equipment or service in question.

Where the amplitude of the received signal is stable, the transition response of the receiving equipment can be maintained at the halfway points on the rise and fall of the voltage waves. The latter then need not be steep, a sine wave or something slightly better being satisfactory for reproduction of the required 50:50 ratio of M/S. For a dot-cycle train, this requires only the fundamental keying frequency. For somewhat more tolerance and better performance on random telegraph keying, practical design often allows for passage of the equivalent of the second harmonic of the fundamental dot-cycle frequency.

Where signal amplitudes vary widely and rapidly, as is common in the case of short-wave signals, the transition response of the receiving equipment effectively slides up and down on the slope of the M-S and S-M transitions in the signal. The M/S ratio then can be held constant at the desired 50:50 value only by provision of a rectangular keying wave shape. This would contain numerous high-order harmonics that would occupy a very wide band and cause interference. A compromise therefore is necessary. Generally speaking, it is satisfactory to hold the M/S ratio within the limits 40:60 to 60:40 at the output of the radio receiver and associated electric circuits, amplifiers, rectifiers, etc. Because of the difficulty of specifying exact wave shapes required in all cases, the Radio Regulations (Atlantic City, 1947) merely stated that for nonfading circuits the third harmonic should be passed, and for fading circuits the fifth. This would require total band widths of, respectively, six and ten times the fundamental keying frequency. Those values today are considered highly conservative and somewhat excessive for commercial operations; the tendency now is to use all possible means for minimizing variations in signal amplitude and to compromise on the third harmonic as the desirable upper limit.

The M/S ratio of the signal, as transmitted or as it arrives at the distant end of a short-wave circuit subject to multipath propagation, may differ appreciably from the ideal 50:50 value. In such cases, either the M or the S intervals in the signal will be elongated and the others shortened. To handle these shortened elements satisfactorily requires a somewhat greater band width in the receiving equipment and associated system. A commonly used approximation has been to calculate an equivalent dot-cycle keying frequency from the length, in milliseconds, of the shortest M or S elements of the incoming signal. This then is used in the usual Fourier analysis.

A more generalized approach to the problem of determining the electrical band width required is based on the permissible slope of the M-S and S-M transitions of the keying wave shape. The method is particularly applicable in cases of mechanized operation, such as teleprinter radio services, where the mechanism of the final utilization device—such as a receiving printer or reperforator—will tolerate no more than a specified maximum of signal elongation or shortening. From this maximum or total there is deducted the value measured or observed on the actual incoming signal. The remainder is the amount of elongation which may be permitted because of the limited band width of the receiving system. By assigning this value to an arbitrarily chosen portion of the transition slope—say from the 10 to the 90 per cent amplitude points on the slope—it is possible to calculate the band width required in i-f filters, etc. The required total (double-side-band) band width, in cps, is approximately equal to 1,000 divided by the number of milliseconds elongation which may be tolerated in the electrical filters, etc., of the receiving equipment and system. For example, if the total permissible elongation is 7.5 millisec, and transmitter maladjustment or multipath propagation introduces 5-millisec elongation, the remaining permissible 2.5 millisec requires a receiver band width of approximately 400 cps. This type of calculation may also be applied to the over-all system comprising transmitter and receiver.

It should be pointed out that any such calculations of required band width, for typical short-wave RT signals, usually involve so many assumptions and approximations that, to the practical man, they must be considered as guides rather than as scientifically exact values. The safe procedure, therefore, is to conduct an experimental check under actual operating conditions.

34. Two-frequency Keying. This method, which sometimes is loosely though erroneously referred to as a form of FSK, is essentially a form of on-off keying in which the mark frequency source is keyed on and the space frequency source simultaneously is keyed off, or vice versa. Side-band and energy distribution, therefore, is that of two carriers being keyed on-off at the speed in use. Since the phase relationship between the two sources generally is uncontrolled, and therefore random, there is a phase discontinuity at transition. This is a possible source of out-of-band

spurious radiation that can cause interference. Furthermore, it is necessary to shape the keying envelope of each M and S suitably to suppress high-order harmonics of the keying frequency that otherwise would cause interference. During the transition interval between such shaped envelopes, the random phase relationship causes either addition or cancellation. This results in a variation in the exact timing of the effective transition. In some cases, at high keying speeds, this may be objectionable.

35. Frequency-shift Keying (FSK).[22–29] This is a form of true f.m., in which the frequency of a single source—an oscillator—is shifted from one frequency to another. For two-condition code this shifting may be from S to M to S or it may be both up and down from a neutral "carrier" frequency midway between the steady M and S frequencies. Since the method is a true f.m., abruptness or smoothness of transition from one frequency to another may be controlled by suitably shaping the wave form of the keying voltage applied to the FSK modulation stage. The resultant FSK signal, being of variable frequency but constant amplitude, then may be passed through class C power amplifiers without affecting the keying envelope shape. Side-band energy distribution thus is determined and controlled by the wave shape of the applied keying and the inherent characteristics of the linear frequency-modulator stage.

36. Bessel-function Analysis. The usual analysis of f.m. is based on the concept of an unmodulated "carrier" midway between the upper and lower shift limits (M and S) of the FSK signal. This unmodulated carrier may be either real or fictitious. In either case, the analysis is the same. Frequency deviation ΔF is reckoned plus and minus from the carrier frequency. The deviation therefore is equal to one-half the value of total frequency shift between M and S, a total shift of 400 cps being a deviation (plus and minus) of 200 cps. Designating the carrier frequency as F, the deviation as ΔF, and the modulating frequency as f, the ratio $\Delta F/f$ is known as the *deviation ratio* or *modulation index*. Side-band distribution is a function of this deviation ratio, the complete theory being given in standard texts on the subject.[27] Data for some typical cases are given in the tabulation below, which is for single-frequency sine-wave modulation.

Deviation ratio $\Delta F/f$	% of unmodulated carrier voltage							
	Carrier	$F \pm f$	$F \pm 2f$	$F \pm 3f$	$F \pm 4f$	$F \pm 5f$	$F \pm 6f$	$F \pm 7f$
0.5	94	24	3.1					
1.0	77	44	11.5	1.96				
2.0	22	58	35	12.9	3.4			
3.0	26	34	49	31	13.2	4.3	1.14	
4.0	40	6.6	36	43	28	13.2	4.9	1.52

More extensive tables of these Bessel functions are available in the literature,[20,27] the usual mathematical terminology for any one coefficient being in the form $J_n(m)$, which designates the Bessel function, of the first kind, of argument m and order n. The argument m is taken as the deviation ratio $\Delta F/f$. The order n, which is the order of the term or side band, is zero for the carrier, 1 for the first pair of side bands at $F \pm f$, 2 for the second pair at $F \pm 2f$, etc.

37. Band Width Occupied for FSK. The above tabulation of FSK side-band amplitudes shows the total band width occupied when an FSK signal is produced by a theoretically perfect f-m modulator, supplied with sine-wave keying at a fundamental keying or dot-cycle frequency f, for various values of frequency deviation ΔF. The full significance can best be brought out by an example. Assume a fundamental keying frequency f of 100 dot-cycles per sec and a total frequency shift, from M to S, of 400 cps. The deviation ΔF then is 200 cps and the deviation ratio is 200:100 or 2.0. The M and S frequencies are at $F \pm 2f$, respectively, and keying side bands of

appreciable amplitude extend to $F \pm 4f$, which is $2f$ beyond the steady M and S frequencies. This is fairly typical of certain present-day commercial operations. At very low values of deviation ratio, below 0.5, the total occupied band width is twice the applied sine-wave keying frequency. At high values of deviation ratio, above 4, side bands of appreciable amplitude extend to values of from $3f$ up to $5f$ outside the steady M and S frequencies.

In RT practice, the wave form of the keying voltage applied to the FS modulator is not sinusoidal but contains harmonics (see Secs. 39 and 40 of Chap. 1) necessary to provide a dot shape that tends more toward the rectangular. Rigorous mathematical analysis of such cases is quite involved and indicates not only resultant components corresponding to the terms of the Fourier expression for the keying wave shape but also components having sum and difference frequencies. The total occupied band width can be approximately determined by considering the value of $\Delta F/f$ not only for the fundamental dot-cycle keying frequency but also, individually, for each of the keying-frequency harmonics in the wave shape of the applied keying voltage. For actual telegraphic keying, where the lengths of M and S intervals range randomly from one to seven unit elements, the energy will not be concentrated on discrete side-band frequencies but will be spread out more at random over the approximate maximum band width indicated by the foregoing analysis for the case of a dot-cycle train.

Side bands extending by undesired amounts beyond the steady M and S frequencies may be suppressed, or reduced, by two general methods. One is to shape the applied d-c keying wave form suitably at the input to the FSK modulator by the use of a low-pass filter. It should be noted, however, that this will have no effect on the production of FSK side bands inherent in f.m. as tabulated in a foregoing paragraph. Such side bands, produced in the FSK modulator or keying stage, can be suppressed or reduced only by the use of a band-pass filter following the modulator stage. This is feasible where the frequency shifting is accomplished at relatively low frequencies, say in the i-f or the v-f range.

38. Band Width Required for FSK. While wire-line practice has arrived at definite answers to the question of band width required for a specified speed of FS keying and type of service,[29] the situation in the short-wave field is not so stabilized. Attempts to obtain international agreement on values of frequency shift, through the International Radio Consultative Committee (Geneva, 1951), resulted in recommendations that necessarily took into account not only technical facts but also operating systems and practices current, at that time, throughout the world. Refinement and simplification of those original recommendations is an evolutionary process that may require some years. Meanwhile, simplified analyses, approximations, and experimental or operating results must serve as a guide to the designer and the practical man.

A total frequency shift of 850 cps was in general use during and following the Second World War when FSK came into widespread use over short-wave circuits. This was used for all keying speeds ranging from slow-speed hand keying to some 85 cps for four-channel TDM. For the latter, this meant an occupied band width of some 1,360 cps. Even a single-channel printer signal, keying at approximately 22 cps, occupied a 1,070-cps band width. Today, with the growing congestion of channels and demand for allocations, these band-width requirements are considered excessive.

Recommended values of total shift (CCIR Geneva, 1951) constitute two series of numbers, *viz.*, 70, 140, 280, 560, 840 cps and 100, 200, 400 cps. Other existing systems employ other values. One such series is 85, 170, 340 cps.

NOTE: CCIR Warsaw, 1956, Recommendation No. 150 specifies preferred values of total shift as 200, 400, and 500 cps; and that values of 140, 280, and 560 may be used provisionally, but that 560 cps should not be adopted for new systems, no mention being made of the 840-cps value.

Maximum utilization of available spectrum space and of channel allocations and the necessity of reducing interference to a minimum indicate the use of the smallest

practicable value of frequency shift for any given service. This might be a deviation ratio of 1.0 or even less. Such small values of total frequency shift, however, run into practical difficulties in the design of transmitters and receivers. Sinusoidal keying also would seem desirable but, particularly at the higher keying speeds such as four-channel TDM, results are more adversely affected by multipath propagation than when the keying wave shape is more nearly rectangular.

While long-term efforts are directed toward the use of minimum values of shift, present practice and trends in commercial application are somewhat more conservative; a reasonable compromise appears to be the use of a deviation ratio of 2.0 at the highest keying speed to be handled by a given radio transmitter or over any one channel thereon. This might, for example, be a total shift of 400 cps for a keying speed of 200 Bauds or 100 dot-cycles per sec. The wave shape of the keying voltage supplied to the FS modulator or keying stage preferably should be shaped by a low-pass filter that will pass no more than the third harmonic of the fundamental keying frequency ($3 \times 100 = 300$ cps).

Twinplex Channeling

Channel A	Channel B	Frequency deviation, cps
Mark	Mark	+600
Mark	Space	+200
Space	Mark	−200
Space	Space	−600

39. Twinplex.[17,18] This is a four-condition mode of FSK (see Sec. 25). The four discrete frequencies, representing the four possible combinations of simultaneous M and S on the two channels, are spaced 400 cps apart, total shift between the highest and lowest thus being 1,200 cps. In a typical application employing a standard 60-wpm teleprinter on each of the two channels, total band width occupied by the frequency shift plus the keying side bands is stated to be approximately 1,700 cps.

KEYING FIDELITY

40. Mark/Space Ratio. Correct and reliable interpretation of received signals, whether by ear or by transcription from an ink-recorder tape record or by machine, requires that the ideal M/S ratio of 50:50 for a dot-cycle be maintained as closely as is practicable. For Morse operations, a range of 60:40 (heavy) to 40:60 (light) is considered generally satisfactory. This is for the entire system and circuit, including all associated transmitting and receiving facilities and the radio path. To maintain this over-all fidelity, it is essential that each link and piece of equipment be appreciably better than the required over-all fidelity, the degree being dependent on the number of contributing elements that make up the whole. While, sometimes, it is possible to correct for heavy or light keying, by compensating adjustments of relays or keying devices in other portions of the over-all circuit, this practice is not recommended because it introduces confusion and resultant delays when facilities must be switched or interchanged. At the slower keying speeds, up to 60 wpm or roughly 25 dot-cycles per sec, the use of electromechanical relays is fairly common and generally satisfactory. While such relays may be used at double that speed, or even higher, the increasing difficulty of maintaining critical adjustments makes such operation generally inadvisable. Present-day practice favors the use of electron-tube, or equivalent, types of keying circuits since these can be designed to provide any required degree of fidelity at even the highest keying speeds.

Checking of the M/S ratio requires suitable monitoring and measuring equipment. For the slower speeds, this may be an ink recorder. At the higher speeds, in particular, a cathode-ray oscilloscope should be used. This has the added advantage of clearly showing the exact wave shape of the keying and all details of minor distortions. The necessity for such a high-speed monitoring and measuring instrument will be understood when it is realized that the entire time duration of a dot, at speeds such as four-channel time-division multiplex keying, may amount to less than 6 millisec (see Fig. 4).

High vs. Very Low Frequencies. The v-l-f services had certain advantages: (1) freedom from interruption due to ionospheric disturbances; (2) 24-hr and year-round operation on a single frequency; and (3) steadiness of signal strengths. Disadvantages were: (1) high cost of transmitting equipment and antennas; (2) high power consumption; (3) relatively low keying speeds; and (4) slowing down or complete stoppage by high levels of atmospheric static, particularly where summer lightning storms were prevalent in the direction of reception. The h-f services also have certain disadvantages: (1) interruption of service by severe ionospheric disturbances, (2) fading, (3) multipath propagation, and (4) necessity for having several frequencies for each circuit to provide 24-hr service the year round and through the sunspot cycle. These disadvantages, however, are far outweighed by the advantages, which include: (1) comparatively low cost, particularly of transmitting equipment and antennas; (2) comparatively low power consumption; (3) high keying-speed capabilities; (4) practicability of highly directive antennas for transmission and for reception; and (5) flexibility of operations.

High Frequency. The outstanding feature of radio propagation in the h-f range, from nominally 3,000 to 30,000 kc, is the relatively low power required for communication over long distances. Typical transoceanic circuits, ranging from 2,000 to 10,000 miles in length, are operated with transmitters rated at from 1 to 50 kw output. Because of the extreme ranges of signal attenuation for such ionospheric or sky-wave propagation,[36] the primary requirement is a proper choice of frequency suitable for the particular circuit (path), time of day, season of the year, and portion of the sunspot cycle. Predicted frequency requirements, both long- and short-term, are issued regularly by government and commercial agencies. Highly directive transmitting antennas concentrate the radiated power in narrow beams, to provide maximum gain and thereby maximum signal strength at the distant receiving station. Receiving antennas preferably are of a directive type providing good signal pickup in the desired direction, and maximum protection against both interference and noise from directions other than the desired one. Diversity reception, preferably space diversity,[37,38] is employed to minimize the effects of short-period fading which normally may cover a 30-db range in amplitude. The use of FSK, instead of on-off keying, provides further improvement[25,39] variously estimated up to 11 db in terms of transmitter power for a given error count in the received signal. Lastly, the question of power rating of the radio transmitter must be decided from considerations of (1) minimum power required to provide the desired type and quality of service over the paths in question, (2) improvement in quality and hours of service obtainable by a given increase in power, and (3) what is economically justified. Required values of received signal strength depend somewhat on noise characteristics of the receiving site and range approximately from 0.1 μv per m for slow-speed aural reception to 5 μv per m for high-speed mechanized operation. Values normally obtained in practice range from effectively zero up to 100 μv per m or even 1,000 μv per m, depending on propagation conditions.

45. Medium-distance Point-to-point Services. In high northern latitudes, where h-f sky-wave propagation is frequently and severely disturbed by auroral phenomena, communication over distances of up to some 1,500 miles often is carried on by means of so-called "surface-wave" transmission at frequencies of approximately 100 to 150 kc. In less disturbed areas, where sky-wave propagation is more reliable, frequencies of from 1,000 to 3,000 kc are employed during the daytime and of from 3,000 to 8,000 kc at night.

46. Short-distance Point-to-point Services. In commercial RT services, the use of frequencies in the v-h-f, u-h-f, and s-h-f ranges is confined chiefly to the remote-control links between central office and the outlying h-f radio stations proper. To handle the required large number of individual telegraph channels over one such carrier, it is necessary to employ some form either of frequency division, such as voice-frequency carrier telegraph techniques, or of pulse-modulation multiplexing.[40] Highly directive antennas, employed for transmission and reception, provide excellent

signal-to-noise ratios for very small amounts of power. Sometimes it is necessary, however, to employ two vertically spaced receiving antennas for space-diversity reception to provide protection against deep fades. At the frequencies generally employed, propagation is essentially line-of-sight and therefore requires the use of repeater stations, ordinarily spaced some 20 to 50 miles apart, the distance being dependent on antenna heights and on the intervening terrain or other obstacles.

Other examples of such line-of-sight operation in the u-h-f and higher-frequency ranges for telegraphic and teleprinter services include radio-relay systems to replace wire-line facilities over heavily loaded trunk routes of domestic telegraph systems[40] and teleprinter communication channels over radio-relay facilities associated with oil pipeline and similar operations in open country or undeveloped areas and over rugged terrain.

47. Mobile Services. Of the three classifications of mobile services, aeronautical, land, and maritime, the oldest and major user of RT communications is the maritime. Commercial services to and between ships employ manual operation except on the largest ships and at major coastal stations where mechanized Morse or printer operation may be available when a large volume of traffic is to be handled and when radio conditions permit. Because of the large number of ships, the limited number of frequency assignments which must be shared, the intermittent nature of the service, and the need to be able to give emergency or distress calls priority, both transmission and reception are handled by, or are under the direct control of, one operator. At major coastal stations, the operator is located at the radio receiver but he remotely controls and keys the transmitters. To permit simultaneous transmission and reception without interference, the transmitters sometimes have been located as far as 50 miles from the main receiving station.

Frequencies generally utilized heretofore for short-distance working by the maritime services lie within the range 350 to 500 kc, the 500-kc frequency being internationally reserved for calling and for distress. A coastal station equipped with such transmitters rated at 5 kw, for example, has a normal daylight range of around 500 miles and a maximum nighttime range of several thousand miles under very favorable conditions on this frequency band. The ranges of shipboard transmitters are somewhat shorter unless working a shore station that is equipped with the best receivers and directional antennas. A frequency of around 150 kc formerly was used by the larger ships for communication over distances of from 1,500 to 3,000 miles under favorable propagation conditions. Present-day practice is to use frequencies in the h-f range (see International Table of Frequency Allocations, Atlantic City, 1947) for long-distance communication in the maritime services. This is because of the great distances that can be covered, with relatively low powered transmitters at these higher frequencies.

The aeronautical services normally employ radiotelegraph communication with airplanes on long overwater flights, teleprinter operation being used for point-to-point communication between fixed stations.

Land mobile service includes two main categories, namely, the Land Transportation Radio Services and the Public Safety Radio Services. These operate on assigned channels in the v-h-f and u-h-f bands and employ radiotelephony rather than radiotelegraphy.

48. Autoalarm.[41] This is an internationally established distress signal for use in the maritime services which is transmitted by hand or by an automatic keying device associated with a main or emergency shipboard transmitter, and which can be received and identified by an unattended autoalarm receiver (see Chap. 19) that will then cause an alarm to sound. The distress signal consists of a series of at least twelve 4-sec dashes separated by 1-sec spaces, transmitted as a tone-modulated c-w emission on 500 kc. American designs of such autoalarm receivers actuate the alarm upon receipt and identification of four such dashes.*

* For special "Alarm Signal for Use on the Maritime Radiotelephony Distress Frequency 2182 kc/s" refer to CCIR Recommendation No. 71, Geneva, 1951.

RADIO TRANSMITTING SYSTEMS

49. Classification. Historically, RT systems and equipment may be classified according to the methods employed to generate the r-f power. Spark transmitters, employing various types of fixed and rotary spark gaps, were once used for all classes of service. Arc transmitters later were employed for long- and medium-distance point-to-point operations and in the marine service. H-f alternators, such as the Alexanderson and the Goldschmidt, were used for v-l-f long-distance point-to-point services as on the international transoceanic circuits. Of these older methods, only the h-f alternator still is in use at a few locations; the others are now obsolete. Descriptions and technical details of such methods and equipment will be found in the older literature.

Tube transmitters, universally used today, may be classified roughly according to type of service (see preceding sections), power output, and method of modulation and of keying. An outline, convenient for the present purpose, is as follows:

- Type of service:
  - Long-distance point-to-point
    - V-l-f
    - H-f
  - Marine
    - Coastal
      - L-f
      - M-f
      - H-f
    - Shipboard
      - L-f
      - M-f
      - H-f
  - U-h-f and microwave relay
    - U-h-f
    - S-h-f
- Power:
  - High
  - Medium
  - Low
- Modulation:
  - Amplitude
    - On-off; class C amplifiers
    - Tone (SSB); linear amplifiers
    - Synthetic SSB
  - Frequency
    - FSK
    - Twinplex
  - Phase
  - Special combinations

50. Power vs. Type of Service. Nominal power-output ratings of radio transmitters, not including power gain of the antennas, may be stated approximately as follows:

Microwave relay	1–5 watts
U-h-f and v-h-f	5–50 watts
Marine:	
Shipboard	100 w–10 kw
Coastal	1–20 kw
Point-to-point:	
H-f	1–50 kw
V-l-f	100–1,000 kw

For the purposes of this chapter it is appropriate to treat radio transmitters from the viewpoint of basic methods. Following sections will therefore be devoted to the various basic methods of modulation and keying and their application to the peculiar requirements of telegraphic service.

51. Amplitude Modulation. *On-Off Keying.* Tube transmitters for this class of service are characterized by the use of nonlinear stages, buffers, frequency doublers and triplers, and class C power amplifiers, following the modulated or keyed low-power stage. Keying may be accomplished by use of an electromechanical relay or preferably, particularly at keying speeds much above 60 wpm or 25 dot-cycles per sec, by the use of a tube keyer. An example of the latter is shown, in circuit form, in Fig. 5. Rounding off square corners of the keying wave shape to reduce its harmonic content may be accomplished by a low-pass filter or an RC circuit ahead of the grid of the d-c keying tube or by use of similar smoothing in the plate supply to this tube. It must be pointed out, however, that the beneficial effects of such smoothing are largely if not completely lost as the r-f signal passes through the

following nonlinear stages that again "square up" the keying envelope of the r-f wave. Therefore, it is customary to place some smoothing or filtering in the d-c plate supply to the final class C power-amplifier stage (for keying analysis and CCIR recommendations, see Sec. 32). However, the inherent voltage regulation of power supplies for plate and bias d-c voltages must be good to prevent the undesired production of triangular or other unsymmetrically shaped keying envelopes of the final r-f output signal.

Fig. 5. On-off tube keyer for transmitter.

Tone SSB. The so-called single-side-band radio transmitter (SSB) is essentially a telephone-type transmitter in which amplitude modulation of the carrier is effected at a suitably low frequency—usually 100 kc or lower—at which it is practicable to filter off and reject one set of side bands and also to suppress the carrier. The resulting SSB signal then is stepped up in frequency by passage through additional modulator stages until the final output frequency is attained. These linear modulators operate at relatively low power levels and are followed by one or more linear amplifiers at the final output frequency. Power output is rated in terms of peak envelope power. This ordinarily is on the order of 2 kw or, with special power amplifiers, may be 20 kw or higher.

For telegraphic or teleprinter services, one voice band of nominally 3 kc is split up into a number of narrower channels by the use of standard v-f carrier-telegraph techniques and filters, the maximum number of such individual channels being determined by the useful band width required on each channel, by the mid-channel spacings, and by the extent to which it is considered permissible to sacrifice signal-to-noise ratio due to division of the total available power between the channels to be used. Practical calculation of this division of the peak envelope power to obtain the level permissible for the tone signal on each channel takes into consideration the small probability that modulation peaks on several or more channels may occur simultaneously. Therefore, it is often satisfactory to make this division not on a peak-voltage basis but rather on a peak-power basis, the peak envelope power being divided by the number of channels. The following tabulation gives both, the 3-, 6-, 9-db values being upper limits on a power-division basis, and the 6-, 12-, 18-db figures being the conservative lower limits on a voltage-division basis.

No. of Tone Telegraph Channels	Channel Level
1	0 db
2	−3 to −6 db
4	−6 to −12 db
8	−9 to −18 db

Keying of the carrier tone for any channel may, of course, be on-off or FSK.

Synthetic SSB.[42] This is a method for producing a high-power SSB signal from conventional designs of double-side-band telephone-type h-f transmitters employing class C r-f amplifiers and high-level amplitude modulation. A suitable adaptor unit, such as the Kahn,[42] generates a true SSB signal at low power level and then separates it into two components—an a-m component and a p-m component. The p-m component is substituted for the output of the conventional frequency source or drive and is amplified by the class C power stages of the transmitter. The a-m component is amplified through the normal voice-frequency amplifiers and high-level

amplitude modulator. With proper relative phasing and proportioning of the two components, the resultant output from the high-level modulator is an exact, but highly amplified, reproduction of the original SSB signal. The method is mentioned here because of its ready adaptability to conventional designs of h-f telegraph or telephone transmitters.

52. Frequency Modulation. Radio transmitters for FSK employ nonlinear class C r-f amplifiers but, unlike those for on-off keying, do not require good voltage regulation of the high-voltage d-c power supplies. This is because the amplitude of the r-f signal

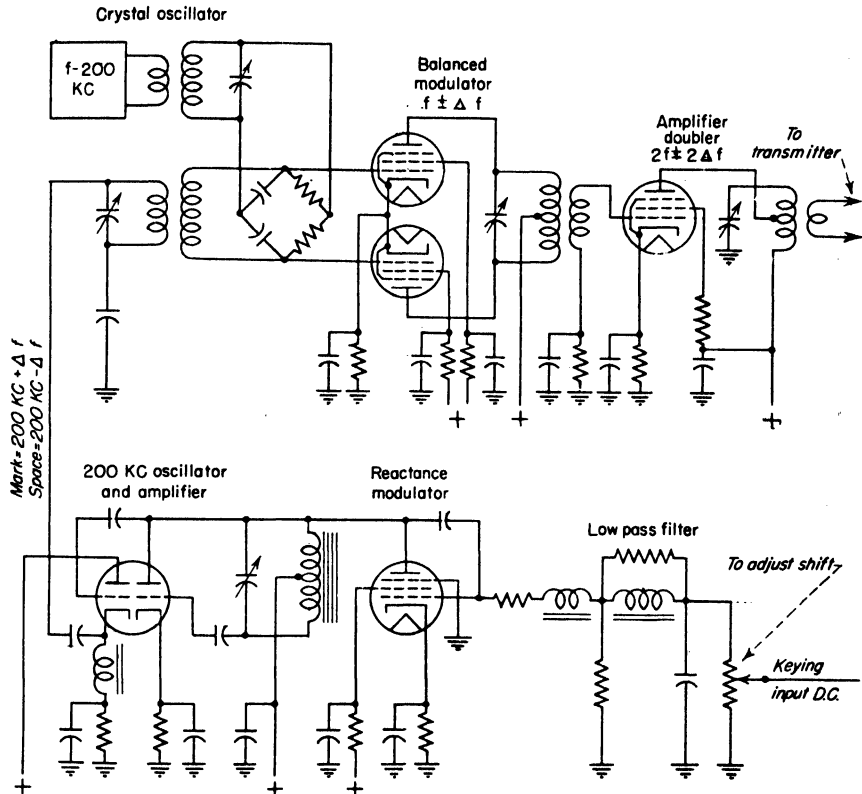



FIG. 6. FSK exciter or drive unit.

is constant, only the frequency changing in accordance with the keying. A commonly used method of generating the desired FSK signal is to replace the conventional crystal-oscillator drive by an FSK exciter or drive unit. Such a unit is shown, schematically, in Fig. 6. The new crystal oscillator, in the FSK exciter, delivers energy at a frequency 200 kc lower than the required submultiple F/n (where n may be 2, 4, or 8, for example) of the final radio frequency. This then is combined, in a balanced modulator stage, with the $200 \pm \Delta F/n$ frequency from a reactance-tube-modulated 200-kc oscillator, the upper set of side bands $(F/n - 200) + (200 \pm \Delta F/n)$ or $F/n \pm \Delta F/n$ being selected and amplified. Frequency multiplication of n times in the multiplier stages of the radio transmitter proper then produces the desired final signal of frequency $F \pm \Delta F$, which is a central carrier frequency F and a total shift, between mark and space, of $2\,\Delta F$.*

* For a discussion of shift values recommended, and resulting band widths, refer to Sec. 38.

Twinplex.[17,18] This system, which is analyzed as regards keying characteristics and band-width requirements in Sec. 39, utilizes an h-f radio transmitter with associated FSK circuits capable of linear shift over a range of something greater than 1,200 cps. With this, there is required a Twinplex combiner which converts and combines the M-S keying of two individual Morse or teleprinter channels into a single four-condition or four-valued d-c signal to drive the linear FSK modulator or keyer stage. Combination is effected by connecting in series the d-c keyed outputs of channels A and B obtained from suitable tone-signal converters. By adjusting the space output of channel A to say 2 volts, and that of channel B to say 1 volt—the mark condition on each being zero volts output—the additive output of the combiner will be as tabulated below. The corresponding values of frequency deviation, in the final output signal, are given in the right-hand column.

Channel		Combined d-c volts	Frequency deviation, cps
A	B		
Mark	Mark	0	+600
Mark	Space	1	+200
Space	Mark	2	−200
Space	Space	3	−600

53. Phase Modulation. The use of approximately 1 radian of phase modulation of the r-f signal at a modulation rate of some 600 cps formerly was quite common on h-f transmitters for on-off keying. This produced a rather wide distribution of the emitted power over a frequency band roughly limited to the channel assignment, the purpose being to provide a form of frequency diversity which could be utilized to reduce the ill effects of fading. The method had value in those cases where reception necessarily was on a single receiver rather than on two or more employed in space diversity. It has been supplanted by modern FSK methods of operation.

Phase shift, instead of frequency shift, has been proposed as a method of transmitting telegraph or teleprinter code signals, the scheme being to shift the phase of a tone frequency, with respect to a fixed reference, by specific discrete amounts to provide one or more channels over a single carrier tone. This would correspond to single-channel and Twinplex modes, respectively, of FSK operation.

54. Special Combinations. Of the various possible combinations of FSK, a.m., on-off keying, and f.m., the best known and most widely used probably is that in which FSK is employed on each of the two or more tones that amplitude-modulate a conventional SSB telephone transmitter.

A less widely used combination is that of the *Twinmode*[17,18] system in which a single r-f carrier is frequency-shifted by one Twinplex (two-channel) keying aggregate and simultaneously is amplitude-modulated by a tone that itself is frequency-shifted by another Twinplex keying aggregate, the particular combination providing four Morse or teleprinter channels over the one r-f carrier.

RADIO RECEIVING SYSTEMS

55. General Characteristics. The broad subject of receiving systems, and radio receiving equipment, is dealt with comprehensively in Chap. 19. The present treatment, therefore, is restricted to those features which are peculiar to, or are of particular importance in, the reception of Morse and teleprinter code signals.

V-l-f reception, at frequencies below 30 kc, is characterized by the use of receivers that will provide the selectivity necessary to separate signals successfully on the 200-cps separation commonly assigned. As a result, permissible keying speeds are relatively low, a practical top figure for this channel spacing being approximately 75 wpm Morse. Another limitation on operating speeds and on the use of teleprinters is the severity of static during the season of the year when lightning storms are prevalent. In the fixed point-to-point v-l-f services, it has been customary to use directive receiving antennas such as the Beverage wave antenna,[43] loops, etc.

The maritime services employ RT receivers which operate in the v-l-f, l-f, and m-f

ranges, covering from 15 to 650 kc. These may be of autodyne or of heterodyne types, designed for convenient tuning and adjustment by the receiving operator, and of emergency and autoalarm types for use on the 500-kc distress frequency. Selectivity may be fixed or adjustable to suit the particular type of service or to provide maximum discrimination against interference.

H-f receivers, to cover the range of 2 or 3 Mc to nominally 30 Mc, vary greatly in details of design to suit the particular service and type of installation. These range from the relatively simple and small receivers of conventional design, suitable for shipboard installations, to the more elaborate types of highly specialized design required by the fixed, point-to-point services. The latter include many features that involve basic design considerations peculiar to the reception of code signals. They will be treated in some detail.

56. H-f Point-to-point Reception. Mechanized operation, particularly five-unit start/stop teleprinter, requires that the receiving system and equipment be designed to have the maximum accuracy and reliability. Since this applies to each unit element (Baudel) in every code combination transmitted, the required accuracy or fidelity of reproduction is, in certain respects, greater than for voice communication. Errors resulting from signal mutilations, due to causes such as noise or fading, should be held to no more than one per thousand words, or 1 in 6,000 characters, for commercially acceptable teleprinter service (see Chap. 27). This requires that not only the average but also the instantaneous signal-to-noise ratio be maintained above the workable minimum, and that the receiving equipment be designed to function and respond properly regardless of wide variations in either average or instantaneous signal level. Directive antennas, employed to minimize interference and noise, may be of such types as the vertical curtain with reflector, the so-called "fishbone" or wave antenna, and the rhombic (see Chap. 20). The receiving station, with its antennas, should be located outside any city and at a distance from radio transmitting stations, high-voltage power lines, and main highways. This is to ensure that local interference and man-made "noise" will be at a minimum.

57. Diversity Reception.[37,38] Instantaneous values of signal strength may normally be expected to vary over a 30-db range in amplitude, such short-period variations or fading having time durations on the order of seconds or fractions of a second—from say 10 sec down to 1 millisec. Observations of h-f signals from distant transmitters show that such fading is not simultaneous as between (1) horizontally and vertically polarized antennas at the same location; (2) two frequencies spaced 500 cps, or less, in the r-f spectrum; and (3) antennas of the same type spaced geographically by distances ranging down to as little as a few wavelengths at the frequency of the signal. This difference or diversity of fading is taken advantage of in the so-called diversity method of reception.

Frequency diversity, which may be obtained by phase- or frequency-modulating the crystal oscillator of the transmitter at a 500-cps rate or by simultaneous transmission of the same keying on two or more modulating tone frequencies applied to the transmitter, is very effective. Generally speaking, the slower the fading the greater the frequency separation required to provide a worthwhile diversity effect. The method requires the use of only one antenna and receiver but is wasteful of frequency-spectrum space.

Space diversity employs two or three antennas spaced approximately 10 wavelengths, or a compromise distance of 1,000 ft, apart, each antenna supplying signal to a separate h-f receiver. The outputs of the individual receivers are combined or switched in such manner as to ensure the best possible signal-to-noise ratio in the final diversity signal that drives a relay or a tone-keyer stage.

Polarization diversity gives the least improvement and so is used only where some form of frequency diversity is impracticable and where space for antennas is very restricted.

Transmitter diversity[44] employs two h-f transmitters radiating from geographically spaced antennas. A very small frequency difference, of some 90 cps, prevents the

setting up of patterns of r-f addition and cancellation between the two radiations and in the receiver. A single receiver of special design is used. The method is of value in special types of service where receivers and receiving-antenna installations must be small and simple, and where the improved reception, perhaps simultaneously at a number of widely scattered points, justifies the increased cost of transmitting facilities.

Improvement obtainable by diversity reception is such that h-f point-to-point RT services consider it essential. To express the improvement in definite figures, however, is difficult because it depends to such an extent not only on antenna or frequency spacing, and on fading conditions, but also on the signal-to-noise ratio. It now is customary to express the improvement in terms of the increase in transmitter power that would be required to give the same reduction in error count as is produced by diversity reception for a given set of conditions and signal-to-noise ratio.[38] For two-receiver diversity, either frequency or space, the improvement has been variously estimated, observed, or calculated as ranging from 5 to 30 db. A representative figure for practical use is 15 db. It is interesting to note that the improvement, expressed in this manner, is greatest for high values of signal-to-noise ratio and least when signal-to-noise ratios are very low. Addition of a third receiver, in diversity combination, provides a further improvement of perhaps 2 to 8 db. While such a rating in decibels is convenient and is of use in over-all system design, the figure of interest to traffic and technical operating personnel is the reduction in error count. At the very low values of signal-to-noise ratio, when mutilations caused by noise "fills" and by fading "drops" are frequent, a "diversity improvement" of only a few decibels may give a 2 or 3 to 1 reduction in error count.

58. Diversity Receivers, General Characteristics.[45,46] In a major point-to-point receiving station that handles a large number of signals simultaneously, the individual h-f receivers used in diversity combination must be of good design as regards all the conventional characteristics and, furthermore, must be well filtered and shielded to prevent r-f or i-f cross talk between receivers and r-f oscillator radiation onto antenna transmission lines and antennas. With transmission lines having negligible stray pickup and an antenna system well balanced with respect to ground, the r-f transfer from balanced antenna system to any unbalanced system, such as coaxial line or a tuned grid circuit, should be through an inductive coupling provided with an electrostatic shield between windings. Otherwise, the polarization and directional characteristics of a good antenna system may be seriously impaired.

Frequency stability of all r-f and i-f heterodyne oscillators is a major consideration where ultimate i-f or a-f band width of a channel may be only 200 cps or less and where the receivers must operate with a minimum of routine attention. Where the use of fixed-frequency crystal oscillators is impracticable or operationally undesirable because of the need to receive on numerous frequencies, oscillators of well-stabilized design and low temperature coefficient are essential. The latter may be supplemented by automatic frequency control, though the use of AFC, on some types of RT signal and under adverse conditions of noise or interference, often is questionable.

Automatic gain control (AGC) is derived from the rectified and combined outputs of the individual receivers; the voltage thus derived is fed back as a bias to the r-f, and sometimes the i-f, amplifier stages of all receivers (see Fig. 7). In diversity reception, this common AGC has a dual role. It maintains the rectified i-f output signal reasonably constant over a wide range of antenna signal strengths; and because it holds the gain of all receivers the same at any instant, the i-f output signal from any one receiver is proportional to the instantaneous r-f signal strength at its antenna. Thus, the antenna having the best signal-to-noise ratio at any instant will contribute most to the output signal. (Qualification of this generalized statement will be taken up later.)

Limiting may be used to ensure holding the level of the i-f or the rectified output signal constant within closer limits than AGC alone will effect, particularly on fading that is too rapid for the AGC to follow. In the case of on-off keying, any such limiting

must follow the final selectivity. It customarily appears in the output relay or the tone-keyer circuit (Fig. 7). Since the envelope shape of the keyed signal supposedly contains at most the fifth, and often no more than the third, harmonic of the fundamental keying frequency, the rise and fall are not instantaneous but have appreciable slopes. Therefore, only comparatively small amounts or degrees of limiting may be used if the M/S ratio of a varying signal is to be reproduced with the required fidelity. In the case of FSK, amplitude limiting does not affect the frequency-shift envelope formation of the keyed wave. Limiting therefore may be, and is, used. This makes it possible to design receivers that are far less critical of adjustment on radiotelegraph

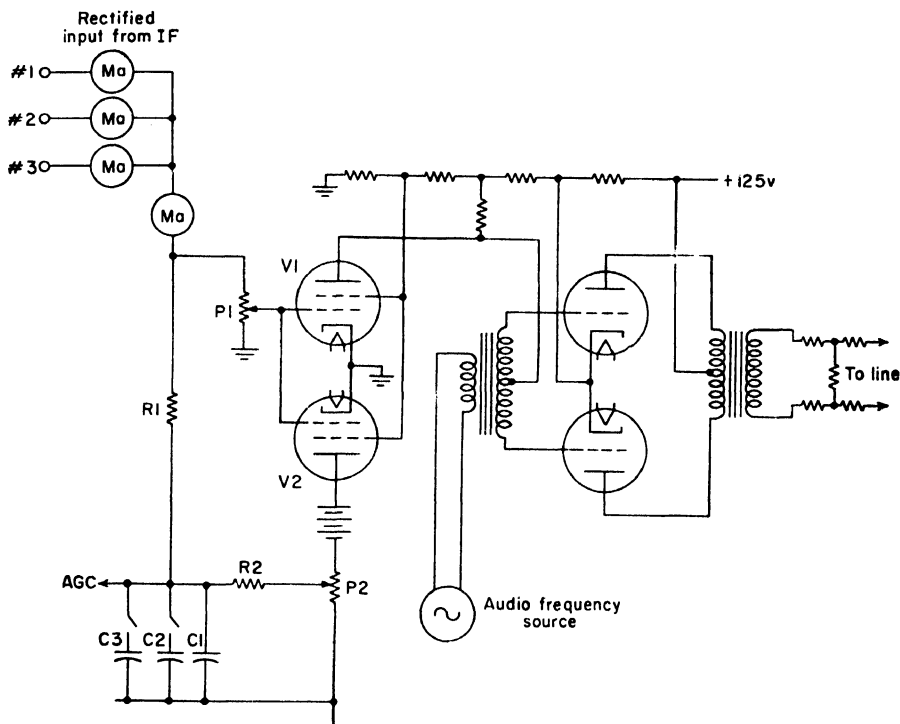



FIG. 7. Diversity combining and on-off tone keyer.

signals which are fading badly. This is one of the outstanding reasons for the practical superiority of FSK over on-off keying for h-f services.

Combining or *switching* of the outputs of the individual receivers used in diversity combination is effected in various ways, the object in all cases being to supply as nearly perfect a signal as possible to the output relay or tone-keyer stage. Typical methods are individually dealt with in following sections.

59. Diversity Receivers, On-Off.[45] Historically and technically, the basic combining circuit for two- or three-receiver diversity reception of a-m or on-off keyed signals is shown by the left-hand portion of Fig. 7. In each receiver, the final i-f signal is rectified and filtered. Filtering is necessary to prevent i-f components from the individual receivers from interacting and producing a-f beats in the combined output when the receivers employ separate rather than common oscillators. Where the i-f system is of relatively low frequency, such as 50 kc, and the modulation components may range up to several kc (voice), it is necessary to use well-balanced full-wave rectification so that the filtering may be made as light as possible, each filter

effectively being capacitively loaded by those of the other receivers paralleled with it. The rectified outputs combine in the common load resistor P_1, from which a suitable voltage is picked off to drive the limiter-reverser tube V_1 which, in turn, changes the bias on the push-pull output stage and thereby keys the tone signal derived from the local a-f source. The final output signal thus is an on-off keyed tone of constant frequency and amplitude.

AGC voltage, fed back to the r-f amplifier stages and sometimes to certain i-f stages of the two or three receivers in use, is derived from the total rectified signal output voltage across the common diode load resistor P_1, R_1 with switchable capacitors C_1, C_2, and C_3 providing a choice of time constant to best handle different speeds and types of fading. A threshold or holding bias to prevent the r-f gain and noise from rising too high during pauses in on-off keying is derived through the coupling-reverser tube V_2 and its plate circuit, the value of this threshold voltage being adjustable by P_2. What this circuit does, in effect, is to deliver an AGC voltage which is derived from the signal when signal is present and from the threshold tube V_2 and potentiometer P_2 when there is no signal present.

Combining the rectified outputs in a common load resistor, as shown in Fig. 7, provides additional switching action over and above that due to the common AGC on all receivers. It is assumed, here, that the signal rectifiers in the individual i-f units are the usual diodes. The voltage across the common load resistor biases these diodes and tends to cut off all but the one which at the moment has the highest value of signal on it. The result is that the rectified outputs of the individual receivers are not proportional to the respective antenna signal strengths, as was stated above when discussing the common AGC. Instead, this difference is accentuated, the receiver having the strongest signal thus supplying most, or all, of the combined output. With individual receiver gains properly adjusted for equal noise level, this inherent voltage regulation or switching action of the common diode-load circuit ensures that the total rectified output will be supplied mostly or entirely by the receiver having the best signal-to-noise ratio at its antenna.

Signal-to-noise ratio in the rectified signal sometimes may be improved by use of a low-pass filter inserted between the point where the three signals are combined and the common load resistor. This method is effective in cases where the band width of the radio receivers is far greater than required for the maximum keying speed, and so long as noise or interference does not drive the i-f systems or the diodes beyond their linear ranges.

The question of two-receiver vs. three-receiver diversity operation has been answered differently by different organizations. For on-off keying, provision of AGC with the threshold feature, and what little limiting is permissible, sometimes is not sufficient to satisfactorily handle adverse conditions of signal-to-noise ratio and fading. At such times, the use of a third receiver appears to be justified.

60. Diversity Receivers, FSK. As regards the r-f and i-f systems of the individual receivers and the manner of deriving a common AGC voltage, diversity receiving equipment for FSK operation is the same as that already described, in Sec. 59, for on-off keying, the AGC voltage for all receivers being obtained through a time-constant circuit connected across the common diode load resistor (see left-hand portion, only, of Fig. 7).

For FSK operation, the diversity receiving equipment differs in two essential respects from that used for on-off keying. These are (1) large amounts of limiting may be used after the i-f selectivity; and (2) the FSK signal must be converted to the on-off form for combining or switching and for driving the output relay or tone keyer (see Fig. 8).

Limiting is used, following the i-f selectivity but before conversion of the FSK signal to on-off, to maintain the i-f output signal and the converted signal at a constant amplitude regardless of individual fades that diversity and AGC have not completely taken care of. The resultant d-c on-off signal thus has a constant marking value, which permits the use of a rounded or even a sinusoidal keying wave shape

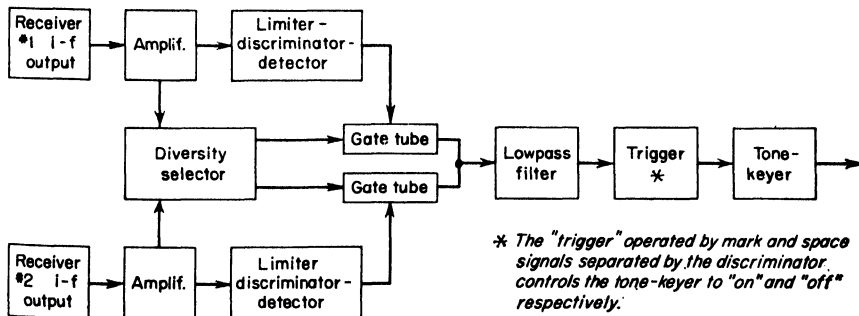
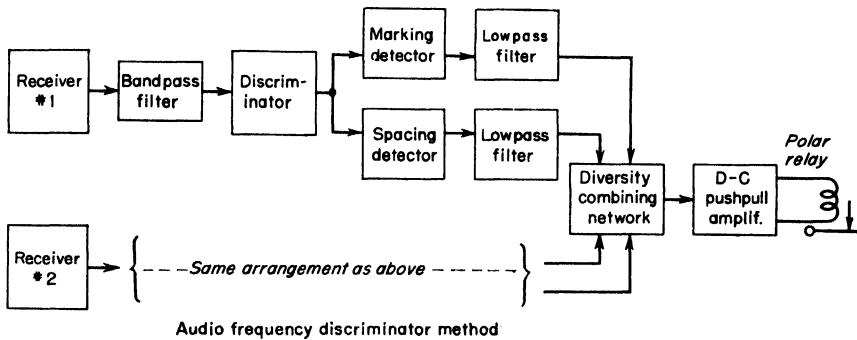
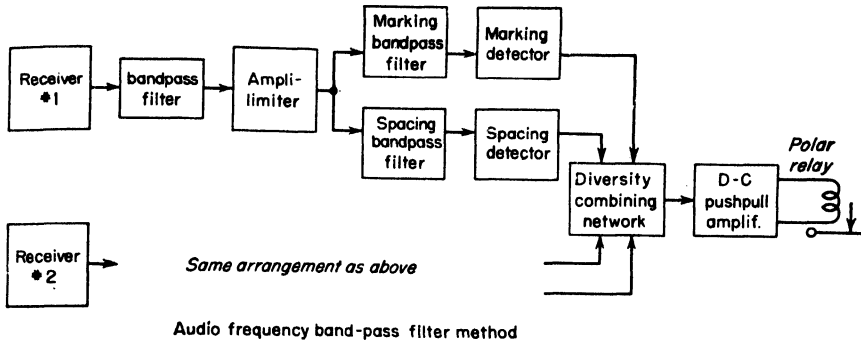



FIG. 8. Receiving methods for frequency-shift keying.

and of setting the operating point, for the following stages and the output relay or tone keyer, at the 50 per cent amplitude level on the rectified keying wave. This ensures that the M/S ratio, in the final output signal from relay or tone keyer, will faithfully reproduce the original keying within close limits regardless of fading.

Conversion of the FSK signal from frequency-shift to on-off is effected by either of the two methods illustrated in Fig. 8 and described below.

Mark and space filters may be used on a signal of the two-frequency-keying type, for which this method is particularly adapted, or on a true FSK signal. In the two-frequency or two-source keying case, the M and S frequencies may have a considerable frequency separation such as the 850 cps used in many applications where standard v-f filter channel spacings were multiples of 170 cps. Whatever the spacing employed, it is preferable to have the filter pass bands centered on the M and S frequencies, respectively, and to have the band widths of each sufficient only for the maximum keying speed to be used. Where the M and S frequencies constitute a single channel, and there are no other channels inserted between them as may be done in tone-channel-multiplex systems, the usual band-pass filters may be replaced by a full-channel-width "roofing" filter and high-pass and low-pass filters for the M and S, respectively. When used on a true FSK signal, the necessary guard band between filters, at the crossover point of their frequency characteristics, imposes a limitation on the minimum usable value of frequency shift. It also is necessary, particularly in the true FSK case, to maintain the central "carrier" frequency of the i-f or a-f signal accurately centered on the crossover frequency between the filters. This may require the use of AFC in the individual receivers, or on common oscillators used to supply the receivers operating in diversity combination. One advantage of the mark-space filter method is that it will hold a steady M or S output for any length of sustained M or S condition on the incoming radio signal.

The so-called *discriminator* method of conversion employs a conventional f-m converter circuit which, for keyed services, may take the form of two resonant i-f circuits tuned to peak at frequencies respectively well above and below the maximum limit M and S frequencies to be received. These two circuits, with their crossing frequency characteristics, supply two detectors or rectifiers designated as the M and the S detectors (see Fig. 8). Permissible values of total frequency shift are limited only by: (1) for minimum shift, by the slope of the discriminator characteristics and the sensitivity of the following detectors; and (2) for maximum shift, by the frequency separation between the resonant or peak frequencies of the two i-f circuits supplying the detectors. A further consideration, however, is the type of coupling from the differential output circuit of the M-S detectors or rectifiers to such following circuits as may be used to further "square up" the rectified signal before it is applied to the output relay or tone keyer. If this coupling is direct—which may involve circuit-design complications—the system is capable of maintaining a sustained M or S output for any such protracted condition of the incoming radio signal. If this coupling is through a capacitor—which simplifies some circuit-design problems—the system will pass the lowest-frequency component of the d-c keying wave but will not maintain steady M or S conditions in the output signal. Some form of mark-restorer, or of space-restorer, circuitry then is required.

Diversity combining or switching, as in the case of on-off equipment, is effected after the signal has been rectified. This eliminates the possibility of r-f or i-f phase-cancellation effects and interreceiver beat frequencies. In the FSK case, however, there are some further considerations. The common AGC, and the voltages to control the gating or switching tubes of Fig. 8, must be derived from the signal after the i-f selectivity but before limiting. The purpose of gate tubes is to provide sharper and more complete diversity switching action than is obtainable with other conventional types of combining circuit.

A low-pass filter, shown in Fig. 8, eliminates undesired modulation components and noise, or M-S beat notes, on frequencies above those required for the maximum keying speed. This filter preferably should be designed to pass the third harmonic of the fundamental keying frequency, as this generally is desirable for consistent maintenance of the proper M/S ratio. However, where keying speeds are low enough not to be appreciably affected by multipath delay, and where the use of limiting and "squaring-up" or "trigger" circuits ahead of the tone keyer permits a sinusoidal keying wave shape at this point, the low-pass filter may be designed to pass only the fundamental of the highest keying frequency to be used.

61. Diversity Receivers, SSB.[46] In a frequency-division tone-channel SSB system, the full benefits of diversity reception can be obtained only by treating each different keying channel as a separate signal and providing it with its own individual, or common-diversity, AGC. This is essential because selective fading makes the individual tone-channel signals fade differently, and therefore an AGC derived from one would not be suitable for use on another unless used in diversity combination. This diversity of fading, as between tone channels, may, however, be used to provide frequency diversity. Whether frequency or space diversity be employed, the common AGC system and combining circuits are basically those already described for on-off and FSK diversity receivers, the method being determined by the mode of keying employed for each telegraph or teleprinter channel.

Each receiver comprises a conventional r-f and i-f system of sufficient band width to pass all required side bands of the SSB signal, and is supplied with its own slow AGC to hold the average signal within reasonable limits and thus prevent overloading. Final selectivity for the individual tone channels is obtained in each receiver by the use of a bank of standard v-f filters or, in special designs, by channel i-f filters at a frequency of some 10 kc.

A typical design provides 16 channels with mid-band frequencies spaced 170 cps, from 425 to 2,975 cps, inclusive. (Note: Other v-f carrier-telegraph systems employ mid-channel spacings of 120 and 150 cps, respectively.) Where one tone channel is to carry the keying of a two- or four-channel TDM aggregate, as is necessary when using present-day designs of automatic-RQ equipment, the band width allotted to this tone channel is a multiple of the basic channel spacing of the system.

On-off keying of a single tone is the simplest application and provides a maximum number of telegraph or teleprinter communication channels for a given number of tone channels. Space diversity must be used, if quality performance is desired.

Two-frequency keying, in which one tone channel is used for mark and another tone channel is used for space—keying on one being the inverse of that on the other—may be employed to provide improved performance similar to that obtainable with true FSK. Early examples employed tones spaced 850 cps apart for M and S.

FSK applied to the carrier tone on each tone channel is the preferred mode of operation.[46] Used with space-diversity reception from two suitably designed receivers and two antennas, this mode provides the maximum in performance and number of communication channels for RT service by FDM over SSB or equivalent facilities.

Frequency diversity may be provided by duplicating each M tone and each S tone, or each FSK channel, on another tone frequency some 500 cps or more removed. This requires twice as much spectrum space or, for a given number of tone channels available, cuts in half the number of communication channels that can be provided.

62. Receiver Selectivity. Final channel selectivity required in the individual receivers used in diversity reception of h-f signals is determined by the speed and mode of keying and the value of total frequency shift in the case of an FSK signal in accordance with keying analyses as given in Secs. 30 and 35. In practice, the optimum band width will be a compromise between that required for high fidelity of keying reproduction and what would be desirable from the viewpoint of minimizing noise and interference. Reasonable compromises, neglecting frequency drift, appear to be as follows.

For on-off keying, the required nominal pass band at, say, -3 db should be six times the highest fundamental keying frequency, a keying speed of 25 dot-cycles per sec, or 50 Bauds, thus requiring a nominal pass band of 150 cps.

For FSK, the required nominal pass band at, say, -3 db can be determined from the tabulation of side-band amplitudes given in Sec. 36, allowing all side bands down to an amplitude of approximately 10 per cent to be passed. It must be remembered that the tabulated values are for a single sinusoidal modulation component and so do not cover third and higher-order harmonic components of the keying wave. The latter may be obtained from the tabulation by assigning suitable values to f (see

Sec. 36). A practical compromise for h-f receiver design purposes is

$$\text{Total nominal band width required} = 2(\Delta F + 1.5f) \quad \text{cps}$$

A flat-topped frequency characteristic, with practically square shoulders, would seem to be the ideal. Such a pass-band characteristic, however, produces an objectionable "overshoot" transient at each M-S and S-M transition of the keying.[47] A somewhat rounded top therefore is preferable.

Skirts of the frequency characteristic, outside the useful pass band, should be as steep as possible or practicable to provide maximum protection against adjacent-channel interference. In practical designs of commercial equipment, the slope of these i-f or over-all selectivity skirts has run as high as 100 or even 200 db per kc (CCIR, Geneva, 1951, Recommendation No. 42 Annex).

Image and other spurious responses, normally arising in heterodyne converters (mixers) or due to overloading by undesired signals of great strength, should be kept below −80 db relative to response to the desired signal. Even greater protection, if economically practicable in the receiver design, is desirable.

REMOTE CONTROL AND SIGNAL TRANSFER

63. General Aspects. Particularly in the fixed point-to-point h-f services, it is necessary to have the so-called central radio office[48] or terminal office located in the business district of a city or close to some major headquarters of operations. The associated h-f transmitting and receiving radio stations often must be located considerable distances from this central control office. In such cases, suitable facilities must be provided for remote keying of the transmitters and for transfer of the received radio signals. If the distance is relatively short, or the number of signals to be handled is relatively small, the simplest and most economical method is the use of d.c. of single or, preferably, plus-and-minus polarity with suitable electromechanical relays. Major centers, however, have their associated radio stations at such distances—now generally within 100 miles—that performance, particularly at high keying speeds, and economy of wire-line or radio-relay facilities dictate the use of standard voice-frequency (tone) carrier-telegraph techniques and equipment. The requisite voice channels are provided over wire-line circuits, or over radio-relay facilities operating in the u-h-f or microwave range.

64. Filter Systems. The basic voice channel over telephone lines or radio-relay facilities has a nominal band width of 3,000 cps though it often extends up to 3,400 or 3,500 cps. This may be divided into two wide-band telegraph channels by the use of a low-pass and a high-pass filter having cutoff frequencies in the neighborhood of 1,500 cps to handle signals of unstable frequency or signals being keyed at high speeds. The more usual arrangement, however, is to employ a bank of band-pass filters to divide the total voice band into a number of channels of widths suitable for the keying speeds to be handled on each tone channel. Typical systems for handling single 60- to 100-wpm printer signals provide 16 or more band-pass channels of 170- or 150- or 120- cps mid-channel spacing as indicated below.

A	B	C
425	525	420
595	675	540
765	825	660
935	975	780
1,105	1,125	900
1,275	1,275	1,020
etc. to 3,145	1,425	etc. to 3,180
	1,575	

A = A.T. & T. Co.
B = Western Union Telegraph Co.
C = CCIT, European, and some United States.

Higher keying speeds, such as two- or four-channel seven-unit TDM, require band widths, and therefore band-pass filters, of greater width. Such filters are designed to pass the equivalent of two, or even as many as five, times the band of a basic channel, being fitted into the over-all plan of channel allocations tabulated above.

Full utilization of high-quality telephone or program circuits, that pass up to some 5,000 or even 7,500 cps, requires the use of special filters. These usually are designed to have relatively wide pass bands to handle high-speed keying such as four-channel TDM. For single-channel speeds, it is customary to utilize two or more standard groups of narrow-band filters by "stacking" one above another in the frequency spectrum. This process of shifting an entire group upward in frequency by, say, 3,000 cps, is a standard technique. At the receiving end of the line, the shifted group is shifted back down to its original frequency range and the individual channels separated out by a bank of standard filters.

Filters at the transmitting end of the circuit need provide only sufficient selectivity to prevent passage of harmonics of the tone carrier frequency and to reduce the amplitude of keying side bands that might fall within the pass band of the adjacent channels.

Filters at the receiving end of the circuit must provide the selectivity required to select the desired channel only and prevent interference from signals on adjacent channels. Such filters may, of course, be used as transmitting filters also.

Optimum performance of filters is obtained only when they are supplied from, and work into, a constant resistance of the value for which the filter is designed. If source or load is of varying or questionable impedance, the filter should be partially isolated from it by insertion of a resistance pad, of proper impedance in both directions, having an attenuation of from 6 to 10 db.

65. Tone Keyers. The carrier tone on each filter channel must be keyed in accordance with the telegraph or teleprinter signal that is to be sent over the channel. On-off keying generally is employed because of its relative simplicity and the system flexibility made possible by its use. While FSK of the channel tones is in wide use on some wire-line systems, and in certain special applications over radio-relay systems, it has not been widely used for remote control and signal transfer in commercial RT systems.

The balanced-modulator type of tone-keyer circuit is illustrated in Fig. 7, where it is driven or keyed by a grid-bias voltage derived from the combined d-c signal out of the diversity receiving equipment. In the case of transmission, where the telegraph keying originates locally in teleprinters or tape transmitters, the latter supply a keyed d-c voltage that swings the grid bias on the balanced-modulator stage from its normal operating or M value to an S value that is well below cutoff for the tubes. The result is that the steady tone signal, from the a-f source of Fig. 7, is respectively amplified and blocked by the push-pull amplifier tubes.

Minimization of keying transients is effected by (1) use of matched tubes to prevent a large unbalanced d-c component in the primary winding of the output transformer, (2) use of maximum permissible tone excitation on the grids at all times to provide a maximum ratio of output tone to residual unbalance transients, and (3) adjustment of final output level by use of an attenuator following the output transformer. An alternative, but less desirable, arrangement is to connect a potentiometer between the two halves of the plate winding of the output transformer and connect the plate supply to the slider on this potentiometer. The latter then is adjusted for minimum keying transients by observing the keyed tone output signal on an oscilloscope.

The output impedance of such a tone keyer at the terminals of the output transformer swings between its nominal value of say 600 ohms on M to a very high value of possibly many thousands of ohms on S. As this constitutes an unsatisfactory termination or source for most filters and lines, an isolating pad of some 6 to 10 db should be permanently associated with the tone keyer.

If the input impedance of the input transformer is high enough for bridging use, a number of such tone-keyer units may be supplied from the same tone source. In a

large installation handling many circuits and signals, this provides both economy and flexibility.

The ring modulator may be used as a tone keyer (see Fig. 9). Diodes may be of the copper-oxide or germanium-crystal types. With such a ring modulator the tone source must supply sufficient voltage and actual power to provide the required output level of keyed tone after a loss of from 6 to 10 db in the ring modulator. If more than one such tone keyer is to be supplied from a common tone source, the latter must be of sufficiently low impedance to supply the several loads with negligible voltage drop or interaction.

66. Tone Levels. The steady mark output from all tone-keyer units, in any one system or installation, should be adjusted to a standard level. This facilitates monitoring and trouble shooting, provides flexibility of operations, minimizes cross talk within or between filter groups, and reduces the possibility of accidental overloading and consequent intermodulation in telephone-line repeaters or radio-relay channels. The standard level, in a 600-ohm circuit, often is established at 6 mw at the monitoring points. It may, however, be considerably lower in some systems, particularly at the input to the telephone line or radio-relay channel, exact values being dependent upon, and specified for, each case.

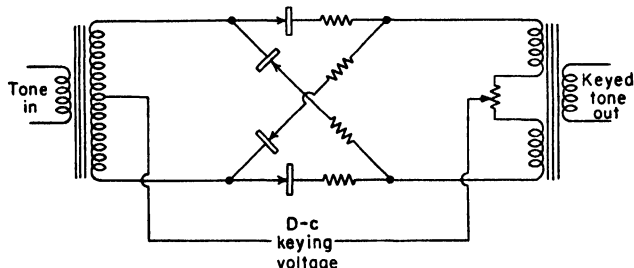



FIG. 9. Ring-modulator tone keyer.

67. Tone-signal Converters. For keying control of the usual RT transmitter, or for actuation of telegraph recorders and teleprinters, the keyed tone signal received over a telephone line or radio-relay channel must be amplified and then converted to an on-off d-c signal of single, or of plus-and-minus, polarity. This is accomplished in a so-called tone-signal converter, which consists of a conventional v-f amplifier of suitable frequency range and gain followed by some form of rectifier. Use of a full-wave rectifier will double the ripple frequency on the rectified signal and thus simplify the problem of designing smoothing circuits or a low-pass filter that will smooth out the ripple with a minimum effect on the keying wave shape. It often is desirable or necessary to maintain the M-S and S-M transitional operating point at a fixed level on the wave shape of the rectified signal regardless of moderate variations in the level of the received tone signal. This is of particular importance where the keying envelope shape of the tone signal received over a narrow filter channel approaches sinusoidal rather than rectangular shape. In such applications, the tone-signal converter is designed with special circuits which hold the transitional operating point at the 50 per cent level within quite close limits over a considerable range of tone-signal input level. This ensures the correct M/S ratio in the rectified output signal regardless of moderate variations in the level of the tone signal received over the filter channel.

TERMINAL-OFFICE EQUIPMENT

68. Morse Equipment. In mechanized or "automatic" Morse operation, a keyboard *perforator* (Fig. 3) is used by the transmitting operator to prepare a perforated tape (Fig. 10) which then is run through a tape transmitter to produce the keyed electrical signal for controlling the radio transmitter. The perforated tape, as illus-

trated in Fig. 10, contains a row of center or feed holes which engage with the teeth of a star wheel, in the tape transmitter, to feed the tape through continuously at a constant speed which is adjustable over a wide range to give traffic keying speeds of from 10 or 15 to several hundred wpm.

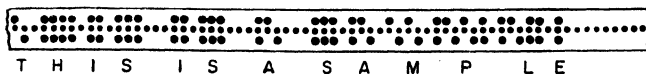



Fig. 10. Morse perforator tape.

The tape transmitter consists essentially of two cam-operated sensing pins or rods that alternately come up against the tape at every point where a perforation might exist, that is, opposite each center hole. If there is a perforation, the sensing pin goes through and thus causes an electrical contact to be closed by a lever or other mechanism associated with the lower end of the sensing pin. If no perforation exists, the sensing pin is stopped by the tough paper tape and the electrical contact is prevented from closing. The two sensing pins which line up with the two rows of perforations in the tape control the "make" or "mark" contacts and the "break" or "space" contacts. The pins alternate in their up-and-down movement against or through the tape and are sufficiently offset in the direction of travel of the tape to produce a Morse dot signal when upper-row and lower-row perforations are opposite the same center hole, and to produce a Morse dash signal when the upper-row and lower-row perforations are opposite adjacent center holes. (In Fig. 10, read the perforations from left to right.)

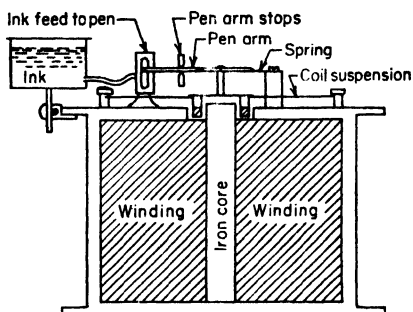



Fig. 11. Ink recorder. Paper tape and tape guide not shown.

Reception, particularly at the higher speeds, is effected by recording the incoming Morse signal on an *ink recorder*, or *undulator*, as shown in Fig. 11. This type consists of a small coil of many turns of very fine wire suspended in a strong radial magnetic field supplied by an electromagnet, the magnetic circuit of the latter consisting of the central core, the iron shell and cover plate, and the air gap in which the small signal coil is suspended. D-c pulses, of the amplified and rectified Morse signal, are sent through the signal coil of the recorder. Marking current causes the suspended coil to move upward and drive the pen arm against the upper stop. Spacing

T H I S I S A S A M P L E

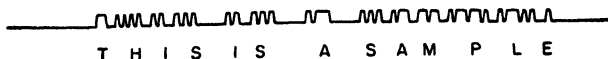



Fig. 12. Sample of Morse ink-recorder tape.

current—zero or of opposite polarity from marking—causes the coil to be returned to the lower stop by action of the spring and the coil suspension, or by the reversed polarity, as the case may be. Mark and space (positive and negative) polarity currents are employed, instead of single polarity on-off, to improve the action of the device at high speeds; such double-current signals are obtained from a pole-changing relay or from a special design of tone-signal converter that gives a rectified output of opposite polarity for mark and space. Speeds attainable run to several hundred wpm.

The tape record is illustrated in Fig. 12.

Printing telegraph machines which operate directly from a Morse-code signal are used on some RT networks, notably the British.

69. Teleprinter Equipment. In international RT services printing telegraph equipment commonly is referred to as teleprinter equipment or by specific trade names as "Teletypewriter." As pointed out in Sec. 16 of this chapter, equipment in general and most widespread use on RT circuits is the standard five-unit start/stop equipment of wire-line services. The code is shown in Fig. 2, a typical keyboard in Fig. 3, and operating speeds in the tabulation of Fig. 4 (see also Chap. 27, Secs. 35, 43).

Cooperation between start/stop teleprinter machines or networks requires that they utilize (1) the same basic code (five- or six-unit); (2) the same code-character assignments; and (3) the same time duration, in milliseconds, of the start and the five code-selecting unit intervals. The stop interval, following the last code-selecting interval, may be of unit-element or any greater duration. This is because the standard receiving mechanism comes to rest at the end of each code combination received after approximately 6.5 unit intervals. Therefore, the duration of the stop interval in the transmitted signal may be anything greater than a nominal minimum of unity. This explains the various total unit intervals per character of 7.0 or 7.42 or 7.5 units tabulated in Fig. 4.

NOTE: An exception to the above is certain types of start/stop signal regenerator which operate on a 7.0-unit cycle, instead of 6.5, and therefore require a transmitted stop interval of greater than 1.0 to supply sufficient margin for operation.

The basic machines are:

Send-receive keyboard printer
Keyboard tape perforator
Tape transmitter-distributor
Receiving printer—tape or page
Reperforator—typing or nontyping

The send-receive keyboard printer is commonly used in customer-to-customer services such as the TEX Overseas Teleprinter Exchange Service, on customer tie lines for filing and delivery of radiotelegrams, and on domestic wire-line circuits to inland points. Other commercial RT services generally employ tape transmission, requiring the use of keyboard perforators and tape transmitter-distributors, and tape relay requiring typing reperforators. Receiving printers generally are of the so-called page type which prints the received message line after line just as an office typewriter does. This eliminates the necessity of pasting tape onto a message blank, as must be done when tape printers are used.

Perforated tape for five-unit teleprinter services has a continuous row of feed holes and five rows of perforations corresponding to the five columns of Fig. 2, where circles indicate tape perforations. The tape transmitter steps this tape along, one feed hole or character at a time. While the tape is at rest, for one feed hole, five sensing pins come up against the tape. Those finding perforations pass through and cause electrical contacts within the tape transmitter to close and provide the M condition for these code elements. Where a sensing pin finds no perforation in the tape, its electrical contact assumes the S condition. In some types of equipment, such as synchronous multiplex equipment, reperforators, etc., the M and S voltages from these five sets of contacts are presented simultaneously and set up the desired combination for further processing. In start/stop systems, however, these five elements must be distributed in time to follow one another in the proper sequence and duration, and they must be preceded by a unit-length start interval and followed by a stop interval of specified duration. These functions are performed by a distributor associated with the tape transmitter proper, the resulting signal for one code combination, in a typical system, having a total duration of $22 + 5(22) + 31 = 163$ millisec.

70. Multiplex Equipment. The TDM mode of operation has been briefly described in Sec. 23. For teleprinter service employing equal-length codes such as the five-, or six-, or seven-unit synchronous or the five-unit start/stop of 7.0 total units there are two basic TDM modes. These are the sequential and the interleaved.

Sequential. The sequential system transmits an entire code combination (one letter, character, or operation) for one channel before starting transmission of a code combination on another channel. This is illustrated below.

Channel I	T H I S
Channel II	W H E N
Channel III	J U N E
Channel IV	F O U R
Aggregate	TWJFHHUOIENUSNER

For the four-channel case illustrated, the time duration of a single-channel unit element is divided into four equal parts. These ¼-length intervals are the unit elements of the four-channel aggregate signal. Thus, four characters—one on each channel—are transmitted in the time normally allotted to the transmission of one character in a single-channel system. At the receiving terminal, a synchronous distributor breaks down the aggregate signal and distributes characters to the respective channel printers.

Mode	Sequence
1	A C B D A C B D A C B D
2	A B A B A B
3	A B A B A B A B A B A B A B C-1 C-2 C-3 C-4 C-5 C-6 C-7 D-1 D-2 D-3 D-4 D-5 D-6 D-7
4	A-1 A-2 A-3 A-4 A-5 A-6 A-7 B-1 B-2 B-3 B-4 B-5 B-6 B-7 C-1 C-2 C-3 C-4 C-5 C-6 C-7 D-1 D-2 D-3 D-4 D-5 D-6 D-7
5	C-1 C-2 C-3 C-4 C-5 C-6 C-7 D-1 D-2 D-3 D-4 D-5 D-6 D-7

FIG. 13. Multiplex channeling modes.

Typical equipment now used on short-wave RT services employs a six-unit code instead of five-unit and provides channel converters to convert from five-unit start/stop (7.42-unit) to six-unit synchronous and, at the receiving end, from six-unit synchronous back to five-unit start/stop (7.42-unit). This permits the use of standard start/stop teleprinter equipment in the terminal offices and interconnection with existing wire-line systems.

Interleaved. The interleaved system also divides each single-channel unit element into shorter aggregate-signal unit elements corresponding in number to the number of channels desired. In a four-channel system, the first, second, third, and fourth quarters of the subdivided single-channel unit element are allocated to the first, second, third, and fourth time-division channels, respectively. Designating these as channels A, B, C, and D, various modes of use are illustrated in Fig. 13.

Channeling. Mode 2 of Fig. 13 is the simplest case, which is that of a two-channel interleaved aggregate. Mode 1 is a four-channel aggregate consisting of two interleaved two-channel aggregates AB and CD.

It should be pointed out that the interleaved mode of TDM is not limited to the use of equal-length teleprinter codes, as is the sequential mode, but permits the use of unequal-length codes, such as Morse, and of different codes on the individual channels.

Mode 3, of Fig. 13, is a four-channel aggregate consisting of a two-channel interleaved aggregate AB interleaved with a two-channel sequential aggregate CD. Mode 4 is a four-channel aggregate formed by interleaving two two-channel sequential aggregates AB and CD.

The sequential two-channel aggregate or subaggregates of Fig. 13 are known as *diplex pairs* and, in the seven-unit form shown, are utilized to handle seven-unit pro-

tected code channels in pairs as required by present designs of Automatic-RQ equipment.[10] The interleaved two- or four-channel subaggregates, or aggregates, may be used for five-unit start/stop (7.0-unit) channels.

Earlier designs of TDM equipment employed rotating commutators with stationary brushes to provide interleaving, and segmented rings on stationary faceplates with rotating brush arms to provide sequential channeling. While such electromechanical equipment still is in use, later designs utilize electronic circuits and techniques to perform the functions and modes of channeling described.

Synchronization. This is a major feature of all TDM systems. The speeds of transmitting and receiving terminals are controlled by frequency standards, such as electrically driven forks, having an accuracy and stability on the order of 1 part in 100,000. In addition, the receiving terminal must be equipped with circuits that automatically correct the phase of the receiving distributors to hold them in exact synchronism and phase with the aggregate keying of the incoming multiplex signal.

Individual channel signals from tape transmitters or from keyboard printers must be synchronized and phased with the transmitting terminal to which they go. The tape transmitter may have its start-up, on each code combination transmitted, controlled by the multiplex terminal. (This is termed *relaxation control.*) Instead, suitably designed converters or adaptors, associated with the multiplex terminal, may be used to accept a nonsynchronized signal and convert it to a properly synchronized and phased signal for presentation to one channel of the multiplex system.

Speeds. The question of speeds of the multiplex equipment and of associated start/stop equipment has not as yet been fully resolved and standardized. Typical systems require either (1) that the multiplex equipment run at a synchronous speed some 2 per cent greater than the operation-per-minute speed of the printer signals supplied to it, or (2) that the tape transmitters, relaxation-controlled by the multiplex terminal, must have a free-running speed some 1 or 2 per cent greater than the synchronous speed of the multiplex equipment.

Receiving printers used on the individual channel signals from the multiplex receiving terminal generally are standard five-unit start/stop machines; or, where a channel must be extended into a domestic wire-line system, a typing reperforator often is used and the tape is then fed through a suitable tape transmitter closely associated with the reperforator. Through the medium of tape storage this arrangement makes it possible to interconnect synchronous multiplex and start/stop channels that operate at different speeds or that use a different total number of unit elements per code combination, such as 5 or 7.0 or 7.42.

71. Automatic-RQ Equipment. The Automatic-RQ mode of operation[10] has been briefly described in Sec. 21. Equipment in use on commercial circuits between

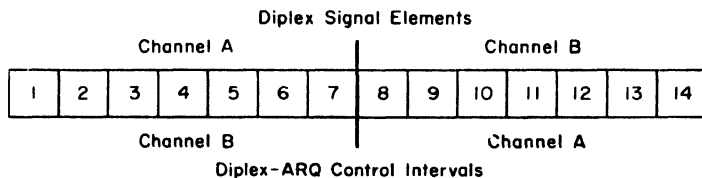



FIG. 14. Sequential diplex time-interval assignments.

many countries of the world is designed to use the seven-unit constant-ratio protected code[6] for detection of mutilated code combinations (see also Sec. 5). Such ARQ channels are arranged in so-called diplex sequential pairs, according to mode 5 of Fig. 13, two such diplex pairs being interleaved, as in mode 4 of Fig. 13, for four-channel multiplex operation with ARQ protection on all channels.

The 14 time intervals of one sequential diplex pair are assigned to the two channels in the manner shown in Fig. 14, the total time duration being that of one code combination, or character, on a single-channel teleprinter. This is nominally $7 \times 20 = 140$ millisec in the European and international system, and $7 \times 22 = 154$

millisec in the United States system of teleprinter operation. (Actual values differ somewhat from these nominal values.) It will be noted that channel A utilizes diplex elements or time intervals 1 to 7, inclusive, for signaling and elements 8 to 14, inclusive, for local control functions required by the synchronous system and the Automatic-RQ equipment. Similarly, channel B utilizes time intervals 8 to 14, inclusive, for signaling and 1 to 7, inclusive, for controls.

With regard to local controls, typical equipment utilizes the time duration of elements 13 to 14 for detecting a mutilated seven-unit code combination, received over channel A, and for initiation of the automatic request for repetition; similarly elements 6 to 7 are used for channel B mutilation detection.

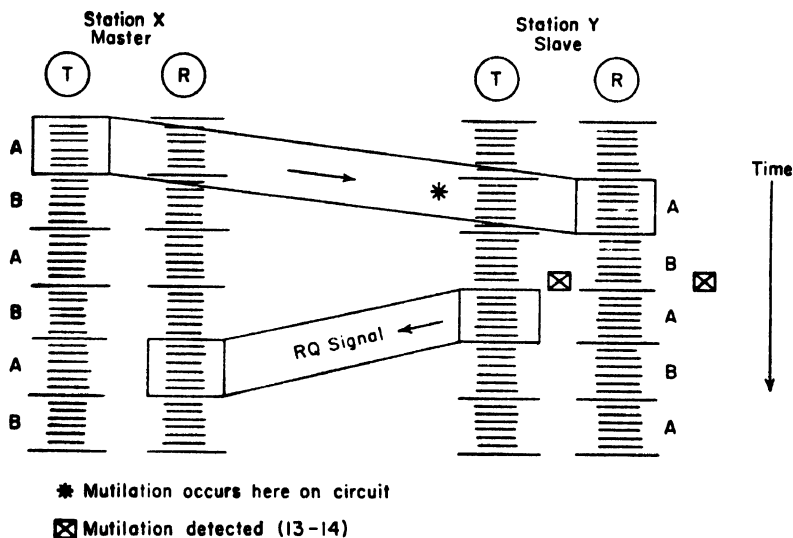



FIG. 15. Automatic-RQ phasing for circuit-transmission time of seven diplex unit elements, and sequence for single mutilation and RQ.

Designs of ARQ equipment presently in use, or under development, range from the basically mechanical to the basically electronic. However, since all must be compatible, all perform the same basic functions. Essentially, these functions are:

1. Synchronism and phasing
2. Code conversion
3. Storage
4. Detection of mutilations
5. Automatic request and repetition

1. *Synchronism and Phasing.* As in any synchronous TDM system, maintenance of exact synchronism and proper phasing is essential. This is effected in the associated multiplex equipment or in the combined MUX-ARQ terminal where the two are designed as a single assembly. In the ARQ case there is an additional phasing requirement. This will best be understood by reference to Fig. 15, which shows a seven-unit code combination on channel A transmitted from Station X and mutilated in the radio circuit before being received at Station Y; detection of the mutilation; and automatic transmission of the RQ signal on channel A from Station Y back to Station X, where the requested repetition will be supplied from storage.

In the simple case depicted in Fig. 15, the total one-way time of transmission over the circuit is equal to the duration of seven diplex elements. Therefore, the receiving (R) and transmitting (T) portions of the equipment at a given station are in phase as

shown. For other values of time of transmission over the circuit—radio and connecting wire lines, etc.—the phase relationship between the R and T sections at the Slave Station Y may have to be adjusted so that the end of an A-channel code combination in the R section does not coincide with the 13-to-14, mutilation-detecting interval of the T section at either station.

Since the phase relationship so established must be maintained, it is essential that all synchronous portions of the T and R equipment at both stations be controlled from a single frequency standard. This latter is located at the Master Station X, where it directly controls the T portion of the local terminal. At the Slave Station Y, the R portion of the terminal automatically keeps itself synchronized with the incoming signal and, in addition, supplies a controlled and phased slave standard frequency to drive the T portion of the slave terminal. In this way the signal received back at R of the Master Station X is maintained in synchronism with, and in proper phase relationship to, the master frequency standard at Station X.

2. *Code Conversion.* As indicated by Figs. 14 and 15, the entire ARQ system is based on use of a constant-ratio protected code. This is the code in which all intelligence is conveyed over the actual radio circuit. However, at the terminal stations or offices, it is desirable to employ standard five-unit start/stop perforators, receiving teleprinters, reperforators, etc. (except for special seven-unit monitor printers used in locating troubles). Therefore, it is necessary to convert from five-unit start/stop code to seven-unit synchronous, and vice versa (for codes, see Fig. 2).

Suitable code converters form a part of the transmitting and the receiving ARQ equipment at each terminal office. These converters may be of various kinds, typical equipment employing a bank of multicontact relays. Other methods of obtaining the required conversion are based on the use of matrix electrical networks or on the use of mechanical selection similar to that employed in standard teleprinter equipment. Code-combination assignments of the ARQ seven-unit code, as given in Fig. 2, have been chosen to facilitate conversion to and from the five-unit code in a particular design of relay-type converter.

3. *Storage.* A sufficient number of code combinations (characters) must be stored at the transmitting terminal where they originate to make available the requested repetition following receipt of the RQ signal after a total delay such as indicated in Fig. 15. The total time of transmission over the radio circuit and its associated wire lines, plus time consumed in equipment functioning, is adequately taken care of by storage of the last three code combinations transmitted. As each code combination of the message is transmitted from the ARQ terminal, one is wiped out from storage and a new one is admitted from the tape transmitter associated with that channel. Thus continuous transmission is provided except when a repetition is requested.

4. *Detection of Mutilations.* In transmission over the radio circuit, a seven-unit code combination may have its proper 3:4 ratio of M/S elements changed by fading, noise, or interference. In the multiplex-ARQ equipment, regeneration of unit intervals provides a signal in which there can exist only a full integral number of M and of S unit elements. A mutilation will therefore result in M/S ratios such as 2:5, 1:6, 4:3, 5:2, or 6:1. Such deviations from the proper 3:4 ratio can readily be detected by a Wheatstone-bridge circuit in which the net resistance in one arm is controlled by a group of signal-operated relays associated with the seven individual unit elements. Balance or imbalance of the bridge—caused by receipt of a valid 3:4 combination or of a mutilation, respectively—may operate another relay which provides the required circuit, polarity, or pulse to control the functioning or blocking of the printing function and, in the latter case, to cause the request for a repetition to be initiated.

Transposing mutilations, in which both an M and an S element are changed within one code combination, result in valid 3:4 combinations other than the one transmitted. These are not detectable as mutilations. Fortunately, the relative frequency of occurrence of such undetectable transpositions is small.[49]

5. *Automatic Request and Repetition.* Detection and registering of the receipt of a mutilated code combination, as described under (4) above, initiates a sequence of functions as follows:

1. At the local (receiving) office Y:
   a. Prevents printing or other functioning of the receiving printer or reperforator by disabling the circuit to it
   b. Stops outbound transmission on that channel by stopping the tape transmitter
   c. Sends out the RQ signal (see Fig. 15)
   d. Resumes outgoing transmission
2. At the distant (transmitting or originating) office X:
   a. Performs no printing or other function on the receiving printer or reperforator, because of the special code combination for the RQ signal (see Fig. 2)
   b. Stops outbound transmission on that channel
   c. Requests and obtains repetition of the mutilated code combination from storage
   d. Resumes normal outbound transmission, with the code combination immediately following the one for which a repetition had been requested
3. At the local (receiving) office Y:
   a. Prints the character, or performs the operation, called for by the repeated code combination if received unmutilated
   b. Resumes normal inbound reception

The case described is the simplest one in which a single mutilation occurs on the circuit in one direction only. More complex cases are ones such as mutilation of two or more consecutive code combinations, mutilations in both directions at the same or nearly the same time, and mutilation of the RQ signal itself. These various possibilities are provided against in the design of the ARQ equipment.

References

1. Telegraph Regulations (Paris Revision, 1949), General Secretariat of the International Telecommunications Union, Geneva, Switzerland.
2. Coggeshall, I. S.: The Transmission of Intelligence in Typescript, *Trans. IRE*, **CS-1** (1), July, 1953; also *Western Union Tech. Rev.*, **7**, January and April, 1953.
3. Hayton, T., C. J. Hughes, and R. L. Saunders: Telegraph Codes and Code Converters, *J. IEE*, **101**, Part III, No. 71, May, 1954.
4. McElroy Chart of Codes and Signals, McElroy Manufacturing Corp., Boston, Mass.
5. Keister, W., A. E. Ritchie, and S. H. Washburn: "The Design of Switching Circuits," Van Nostrand.
6. Moore, J. B., and R. E. Mathes: "Printing Telegraph System," U.S. Patent 2183147 and Re 23028.
7. Hamming, R. W.: Error Detecting and Error Correcting Codes, *Bell System Tech. J.*, **29**, April, 1950.
8. Silverman, R. A., and M. Balser: Coding for Constant-data-rate Systems, Part I. A New Error-correcting Code, *Proc. IRE*, **42** (9), September, 1954.
9. Carter, R. O., and L. K. Wheeler: Code Convertors for the Interconnection of Morse and Teleprinter Systems, *J. IEE*, **101**, Part III, No. 71, May, 1954.
10. van Duuren, H. C. A.: "Printing Telegraph Systems," U.S. Patent 2313980.
11. Sparks, Sidney: International Radiotelegraph Communications, AIEE paper 48-131 presented at Mexico City general meeting, June, 1948.
12. Moore, J. B.: Accuracy and Speed on Short-wave Teleprinter Services, *Proc. National Electronics Conference*, **9**, 927–934, February, 1954.
13. Sparks, S., and R. G. Kreer: Tape Relay System for Radio-telegraph Operation, *RCA Rev.*, **8** (3), September, 1947.
14. Mathes, R. E., J. L. Callahan, and A. Kahn: Time-division Multiplex in Radiotelegraphic Practice, *Proc. IRE*, **26** (1), January, 1938.
15. Hansen, T. A., and R. D. Slayton: An Electronic Time Division Multiplex Telegraph Set, *Trans. AIEE*, **70**, 354–361, Part I, 1951.
16. Hoover, Ray: The Application of Western Union Multiplex to Navy Radio, *Western Union Tech. Rev.*, **2** (1), January, 1948.
17. Buff, C.: Twinplex and Twinmode Radiotelegraph Systems, *Elec. Commun.*, **29**, March, 1952.
18. Buff, C.: Twinplex, Twinmode, and Polyplex Radiotelegraph Systems, *Trans. IRE*, **CS-1**, July, 1953.
19. Shea, T. E.: "Transmission Networks and Wave Filters," Part III, Composition of Transient Waves, Van Nostrand, 1929.
20. "Reference Data for Radio Engineers," 4th ed., Federal Telephone and Radio Co.
21. Bandwidths of Emission, Recommendation No. 36 (Geneva, 1951) of the International Radio Consultative Committee.
22. Carson, John R.: Notes on the Theory of Modulation, *Proc. IRE*, **10** (1), February, 1922.
23. van der Pol, Balth: Frequency Modulation, *Proc. IRE*, **18** (7), July, 1930.

24. Crosby, M. G.: Carrier and Side-frequency Relations with Multitone Frequency or Phase Modulation, *RCA Rev.*, July, 1938.
25. Peterson, H. O., J. B. Atwood, H. E. Goldstine, G. E. Hansell, and R. E. Schock: Observations and Comparisons on Radio Telegraph Signaling by Frequency Shift and On-Off Keying, *RCA Rev.*, March, 1946.
26. Wickizer, G. S.: Relative Amplitude of Side Frequencies in ON/OFF and Frequency Shift Keying, *RCA Rev.*, March, 1947.
27. Hund, August: "Frequency Modulation," McGraw-Hill, 1942.
28. Davey, J. R., and A. L. Matte: Frequency Shift Telegraphy—Radio and Wire Applications, *Bell System Tech. J.*, **27** (2), April, 1948.
29. Bramhall, F. B., and J. E. Boughtwood: Frequency-modulated Carrier Telegraph System, *Trans. AIEE*, **61**, 36, 1942.
30. Rea, W. T.: Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers, *Bell. System Tech. J.*, **23** (3), July, 1944.
31. Cory, S. I.: Telegraph Transmission Coefficients, *Wire and Radio Communications*, May, 1955.
32. Cannon, W. D.: A Portable Telegraph Bias and Distortion Measuring Instrument, *Western Union Tech. Rev.*, **9** (2), April, 1955.
33. Nyquist, H., R. B. Shanck, and S. I. Cory: Measurement of Telegraph Transmission, *Trans. AIEE*, March, 1927.
34. Radio Regulations annexed to International Telecommunication Convention, Atlantic City, 1947.
35. Espenschied, Lloyd, C. N. Anderson, and Austin Bailey: Transatlantic Radio Telephone Transmission, *Proc. IRE*, February, 1926.
36. Ionospheric Radio Propagation, *Nat. Bur. Standards Circ.* 462.
37. Beverage, H. H., and H. O. Peterson: Diversity Receiving System of RCA Communications, Inc., for Radiotelegraphy, *Proc. IRE*, April, 1931.
38. Montgomery, G. F.: Message Error in Diversity Frequency-shift Reception, *Proc. IRE*, **42** (7), July, 1954.
39. Sprague, R. M.: Frequency Shift Radiotelegraph and Teletype System, *Electronics*, November, 1944.
40. Bramhall, F. B.: Radio Relays for Telegraphy, *Elec. Eng.*, November, 1946.
41. Alarm Signal, Sec. IX, Pars. 920–931, of Radio Regulations, Atlantic City, 1947. Also Auto Alarms, Regulation 11, International Conference on Safety of Life at Sea, London, 1948 (U.S. Department of State).
42. Kahn, L. R.: Single-sideband Transmission by Envelope Elimination and Restoration, *Proc. IRE*, **40** (7), July, 1952.
43. Beverage, Rice, and Kellogg: The Wave Antenna, *Trans. AIEE*, **215**, March, 1923.
44. Hansell, G. E.: Transmitter Diversity Applied to Machine Telegraph Radio Circuits, *RCA Rev.*, **11** (4), December, 1950.
45. J. B. Moore, Recent Developments in Diversity Receiving Equipment, *RCA Rev.*, **2** (1), July, 1937.
46. Boughtwood, J. E.: Telegraph Terminal AN/FGC-29, Circuit Design Aspects, *Trans. IRE*, **CS-2** (3), November, 1954.
47. Valley, G. E., and H. Wallman: "Vacuum Tube Amplifiers," M.I.T. Radiation Laboratory Series, Vol. 18, McGraw-Hill, 1948.
48. Rau, D. S., and V. H. Brown: A Modern Radiotelegraph Control Center, *RCA Rev.*, **4** (1), July, 1939.
49. Moore, J. B.: Signal Mutilation and Error Prediction on Short-wave Radio Teleprinter Services, IRE 1957 National Convention Record, Part 8, pp. 127–131.

CHAPTER 25

ELECTRONICS IN AVIATION

By S. B. Poritzky[1]

The tremendous growth of aviation has been inexorably tied to the rapid strides made in the art and science of electronics. Faster and better aircraft have spurred and demanded equally startling advances in electronic equipment. Powerful and complex manned aircraft demand an eye faster than the human eye and response faster than human response to control and use them efficiently. The need for simultaneous operation of many slow and very fast aircraft within a specific geographical area and the efficient and safe handling of these aircraft with respect to each other involve control processes beyond the capabilities of human beings alone. Aircraft are deployed to seek out and destroy an enemy far beyond human sight and traveling so fast and in such intricate maneuvers as to require more than human computation to seek him out. Pilotless aircraft and missiles are used to perform tasks which previously challenged the pilot, in an attempt to reach out tremendous distances and descend on remote targets, meanwhile performing protective maneuvers to elude enemy weapons. In each of these endeavors and the varied others to which the products of aviation are assigned, electronics provides the brains, eyes, ears, and the capability of control. During this phenomenal growth, electronics has graduated from its rudimentary role of providing merely simple communications and entered virtually every phase of aviation.

THE ROLE OF ELECTRONICS

The contributions of electronic equipment in the operation and control of modern aircraft may be subdivided into several categories:

1. The sensors, computers, controls, and displays used in the actual operation of aircraft
2. Devices which are not directly concerned with the actual flying of the aircraft but are tools in the performance of the aircraft mission
3. The communications devices which link the aircraft to other aircraft and to ground facilities
4. The navigation aids which serve as the road signs and highway white lines to aircraft, providing three-dimensional track guidance to the desired objective
5. The traffic-control facilities which enable ground personnel to assure the safe and expeditious movement of many aircraft operating in close proximity to each other

Two yardsticks will help judge the role of electronics in aviation. The electronic equipment installed on the latest civil air carrier aircraft costs more than the full price of a DC-3 aircraft when new. The number of aircraft arrivals and departures at airports having CAA control towers in the United States and its territories in 1946 was slightly over 4 million operations. In 1957, it was more than 16 million operations. In 1960 it is estimated that there will be 22 million such operations.

1. Electronics in Military Aviation. Many of the electronic devices and aids found on modern aircraft have evolved directly from the surge of development of military devices during and since the Second World War. The military designer, searching for the last bit of aircraft performance to achieve a superior weapon, must make heavy demands on electronic devices. In a military aircraft traveling at tremendous speed, in which the pilot must perform many functions, all manner of aids are required to relieve the pilot of routine jobs which machines can do, to free him for his basic

[1] Air Transport Association of America, Washington, D.C.

decision-making responsibility. In addition to the functions of communications, navigation, and traffic control, there must be aids to simplify the flying of the aircraft and to perfect the jobs of fire control, electronic identification, countermeasures, and so on. Although many of the specific problems of the military aircraft are similar in nature to those encountered in civil operations, they are almost always aggravated by the operational need for ever-greater accuracy, greater load capacity, and the capability of operating with less direct and detailed control from the pilot or operator than might be required in a civil situation.

Along with these more stringent requirements go the handicaps which require military equipment to operate over greater extremes of temperature, altitude, shock, and vibration, along with the tremendous handicap that each pound of electronics and each cubic inch of volume used may rob from an aircraft performance which might save the crew's lives.

Development of military electronics must, by its nature, always be in a hurry. In a continuing race for weapons superiority, electronic systems as well as aircraft constantly face the imminence of obsolescence perhaps even before the equipment comes off the production line. Particularly in times of emergency military electronic equipment often cannot be given the long and thorough tests which might be desirable to assure a known-reliable operating system. Electronics in military aviation must frequently take the lead in the exploration of new, untried techniques not only of circuitry but also of operational systems.

In spite of the major differences between the applications of electronics in military and civil aviation, the basic principles of equipment in use are very similar. Although certain special devices, such as electronic countermeasures and fire-control systems, are unique to military operations, the majority of devices which fall into the five categories outlined above perform essentially similar functions in civil and military aircraft. Thus, in surveying the field of aviation electronics, it is appropriate to discuss primarily the many systems and equipments which serve the mutual needs of all aviation.

The basic operational requirements for an aircraft are, first, to travel safely from point *A* to point *B* without interfering with other aircraft performing similar or dissimilar missions and, second, to perform its mission, whether to deliver passengers, cargo, or bombs, to undertake reconnaissance, or to attack another aircraft. The five categories of tools, the sensors, the tools required for the mission, the communications, navigation, and traffic control, are all involved in any aircraft flight. Only the degree and extent of use of the various tools are controlled by the type of aircraft and the type of mission to be performed.

2. Flying the Airplane. In the basic function of flying the aircraft electronics, electromechanics, and the science of mechanics merge to simplify the pilot's job.

Early aircraft, and even some modern small aircraft, not fitted with sophisticated instruments and automatic aids, have employed the senses of the pilot as the basic source of information in flying the airplane.

Several steps are involved in manual operation of an aircraft. The first of these is the "selector" function in which the pilot utilizes his mental processes to decide his destination, route, and flight altitude. Next is the "sensor" function in which the pilot employs his sight, hearing, touch, and balance as sensors to detect deviations from his route as well as such references as altitude, attitude, wind speed, and resultant drift. The "sensor" and "selector" functions are used alternately, since the information developed by the "sensors" usually requires action or decisions to be made by the "selector" function. After selection and sensing, the next step is to convert the decision to act into action. This mechanism of conversion constitutes the "coupler" which converts the sensed deviations and the resultant selected actions into the motor skills of the pilot's hands and feet; thus, the hands and feet serve as the "actuators" to control the rudder pedals and the "stick," which in turn alter the position or path of the aircraft into the selected condition. The fact that the aircraft responds to this chain of selection-sensing-coupling-control and actuation, in effect, tells the

pilot that the aircraft is behaving and proceeding as he wishes. In this fashion, the aircraft proceeds until a new decision is made by the pilot or until his senses tell him that a new decision is required. This chain of events represents a crude form of servomechanism with the human being doing both the sensing of error (deviation from a desired flight path) and the necessary proportional correction or actuation (the application of a command response to an error signal) to restore the aircraft to a desired flight path.

When an aircraft is surrounded by impenetrable fog, the pilot's senses alone are inadequate.

In modern, highly instrumented aircraft, the basic sequence of events described above still exists, but except in the smallest and lightest of aircraft, the methods by which the several acts are carried out are different. A number of electronic and electromechanical aids are available to simplify and improve the control of the aircraft.

3. Substitutes for the Senses. When the weather is bad or when the performance of the aircraft is very high, the pilot can no longer depend upon his senses alone to guide the aircraft. A number of substitutes for the senses have been developed and are in use. Aircraft-altitude information is provided by a simple aneroid barometer calibrated in the altitude equivalent to atmospheric pressure. Air speed, the speed with which the aircraft is moving with respect to the air mass, is obtained by measuring differential air pressures exerted on the aircraft. These instruments and the usual engine instruments in conventional propeller aircraft are primarily mechanical in nature. A more complex sense substitute is a device through which the pilot can be provided with a visual indication of the attitude of his aircraft relative to the horizon. When the weather is good, the alignment of the wings to the horizon serves to provide this information, but this datum is not available when the horizon is not visible. As a substitute for this all-important horizon line, the pilot utilizes a gyroscopic instrument known as an artificial horizon. Although the artificial-horizon instrument is electromechanical, rather than electronic, different applications of the same principle will be met in ensuing portions of this chapter, so that a brief description of the artificial horizon and its major element, the gyroscope, is in order.

4. The Gyroscope. A gyroscope is a spinning wheel mounted in such a way that it can assume any position in space. As shown in Fig. 1, the wheel or rotor is suspended in three sets of gimbals. The rotor spinning around the axis XX is free to turn on that axis on bearings. The gimbal supporting the rotor is free to turn about axis YY on pivots in the outer gimbal. The outer gimbal is free to turn about the axis ZZ on pivots in this portion; thus, the wheel or rotor is free to spin in three degrees of freedom. However, the wheel does not become a gyroscope or exhibit the properties in which we are interested until the wheel spins.

Fig. 1. Elements of a gyroscope.

When the wheel spins, its axis of rotation tends to remain in a fixed direction in space. The wheel resists any force which attempts to change the direction of its axis. Since the axis is at right angles to the plane of rotation of the rotor wheel, the wheel resists any change of its plane. This gyroscopic "rigidity" is maintained with respect to space and not with respect to the earth.

Although the spinning-wheel gyroscope will resist a change of direction of its spin axis, it will move in response to such a force or pressure. This movement is not in

direct response to the force, but the gyro axis will be displaced in a direction at right angles to the applied force, rather than in the direction of the applied force, and in such a way as to tend to cause the direction of rotation of the rotor to assume the direction of the torque resulting from the applied force. For example, if a force is applied in an arc about the axis *YY*, the force is opposed by the resistance of the gyroscopic inertia which prevents the rotor from being displaced about the axis *YY*. If the rotor is spinning clockwise, the force applied about the axis *YY* will tend to cause precession 90 deg ahead in the direction of rotation. Thus, application of the force about the *YY* axis will cause a rotor turn above the *ZZ* axis, and this is a useful phenomenon.

These two properties are basic and vital to the use of the gyroscope as an instrument of flight. Gyroscopic inertia or "rigidity" establishes a reference in space which is free of effect from its supporting body. The method of *precession* described becomes useful in referencing the gyroscope to the earth rather than to its position in space. If a gyroscope is placed anywhere on the earth with its starting position such that the spin axis is horizontal and the gyro is observed from a point in space, the spin axis will remain fixed. Since an observer standing on the earth is, however, subject to the rotation of the earth, the gyro will appear to tilt gradually or drift from the horizontal position. At the end of 6 hr, when the earth has turned one quarter of a turn, the spin axis has appeared to tilt 90 deg and is now in a vertical position. After 12 hr the spin axis is again horizontal but pointing in an opposite direction from the starting position. This phenomenon in a free-spinning wheel or gyroscope is known as *apparent drift*. The spinning wheel will also be subject to friction and slight unbalances and these will also produce forces resulting in gyroscope drift. Since the intention is to provide a wheel which maintains its position with respect to the surface of the earth, corrections must be made for the friction and unbalance and the effect of the apparent drift due to the rotation of the earth.

5. Artificial-horizon Indicator. In this gyroscopic instrument a force to return the spin axis to its normal horizontal position is provided by a mass coupled mechanically to the gyro housing. Gravity controls the erecting mass so that, when the gyro is in its normal horizontal position, the torques about its axis are equal and no force is exerted to precess the gyro. When the gyro drifts from its horizontal position, torque is applied to the gyroscope to cause precession. This precession erects the spin axis of the gyro, and because of the mechanical coupling between the gyro and the mass, the mass is brought back into gravity equilibrium.

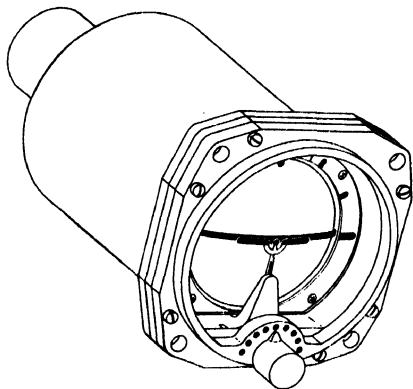



Fig. 2. Gyro horizon indicator.

If the spinning-wheel gyro is housed in a case but given three degrees of freedom with respect to the case and thus with respect to the aircraft, the aircraft and instrument case will move with respect to the gyro when the aircraft attitude changes about either the roll or pitch axis. By providing suitable markings on the gyro and on the instrument-case window through which the gyro is seen, an artificial-horizon line is attained, as shown in Fig. 2. If a horizon-equivalent line is shown on the gyro itself, and a miniature airplane is superimposed on it but fastened to the instrument case (thus representing the aircraft), pitch motion "up" will raise the miniature aircraft "up" above the horizon line, which is maintained steady by the gyro. Similarly, if the aircraft rolls to the right, the right wing of the miniature airplane on the instrument case will roll downward to the right with respect to the horizon line.

6. Gyroscopic Compass (Foucault's Gyroscope). An application of the gyroscope quite similar to the artificial horizon is that of the gyroscopic compass. Here, the

rapidly rotating wheel is kept horizontal by means of a pendulum bob attached to the spin axis. Since the earth is spinning on an axis determined by the geographic north and south poles and the spin axis of a gyroscope is subject to the action of the pendulum, the spin axis oscillates about a meridian with the torque of restitution to the horizontal being maximum at the equator and falling off as the cosine of the latitude, going to zero at the poles. Because of the damping caused by friction and other sources in the gyroscope, the gyro spin axis ultimately comes to rest pointing toward the true geographic north. The gyroscopic compass is independent of magnetic flux and can be used in areas where magnetic disturbances do not permit the use of a magnetic compass.

7. Magnetically Slaved Gyro Compass. Since the magnetic pole, and thus the magnetic meridian, do not align with the geographic poles and the geographic meridian, the directional gyro must be provided with a magnetic sense if it is desired to use the magnetic north as the aircraft-heading reference. For this type of application, the gyro must be slaved to the earth's magnetic field. To accomplish this, the sensing elements for the magnetic lines of force are related to the gyroscope output signal. The signals from the gyro will now be used in a direction indicator only after they are modified by the output of a sensing detector of the magnetic lines of force. The gyro output and the magnetic flux detector are electrically combined in a synchro system and the drift of the gyroscope is controlled so that the gyro spin axis maintains magnetic north as its reference position.

It may be seen that the gyroscope in several forms is an extremely useful tool in substituting for the senses of the pilot. The artificial horizon and the gyrocompass described above serve only as examples of many applications of the gyroscope. The gyrocompass, and particularly the gyrocompass slaved to the magnetic lines of force, constitutes a basic aircraft master direction indicator and has important application in the transition to automatic flight.

8. Sensing and Repeating of Information. The pilot who "flew by the seat of his pants" utilized his physical senses to detect deviation from his route, attitude references, and the effects of his manipulation of the controls. As the process of flight becomes more complex and we begin to depend on aids other than the pilot's senses, the output from such things as gyroscopes and navigation devices can no longer be observed and acted on directly—particularly as the transition to automatic flight begins.

9. Synchro Systems. The need for the display of information at locations other than those where the information is generated, such as the position of control surfaces, gyro orientation, etc., as well as a requirement to exert a force at a distance which may be proportional to but much larger than the controlling force, has resulted in wide usage of synchro systems and servomechanisms in aircraft.

The simplest form of synchro system used in modern aircraft is the transmitter-receiver (repeater) system. In this synchro system, the two units are electrically identical. The transmitter is turned by the original data source and the receiver follows to indicate corresponding angular data at a remote point, as shown in Fig. 3. The voltages across the wires X, Y, Z[1] are single-phase voltages derived by straight transformer coupling from the field produced by the rotor. In a consideration of the potentials existing between the three line terminals of the synchro system the distribution of voltages across the X, Y, Z terminals is illustrated in Fig. 3. The simple transmitter-receiver synchro system operates on the basis of exact alignment between transmitter and receiver rotor-shaft positions. Under the condition of alignment, the line current flowing through the rotor coils of the transmitter and the receiver units induces voltages in the stator coils to these units which are equal and opposite, resulting in a theoretical condition of zero current in the X, Y, Z wires at balance.

When the transmitter shaft is displaced relative to the receiver shaft, the X, Y, Z voltages from the transmitter are no longer equal to those induced in the receiver stator. As a result, relatively large circulating currents flow in the wires X, Y, Z.

[1] Standard ARINC Synchro System Manual notation.

These circulating currents result in the production of a field in the magnetic structure of the receiver. This field, being angularly aligned with the shaft position of the receiver and the now displaced rotor shaft, results in a torque which restores the receiver rotor shaft to alignment with the transmitted angle. This is the simplest form of repeating position information. Repeaters of this sort are used to report many different types of information from control surfaces and from air-borne electronic equipment to the pilot and, in more sophisticated forms, to automatic flight equipment.

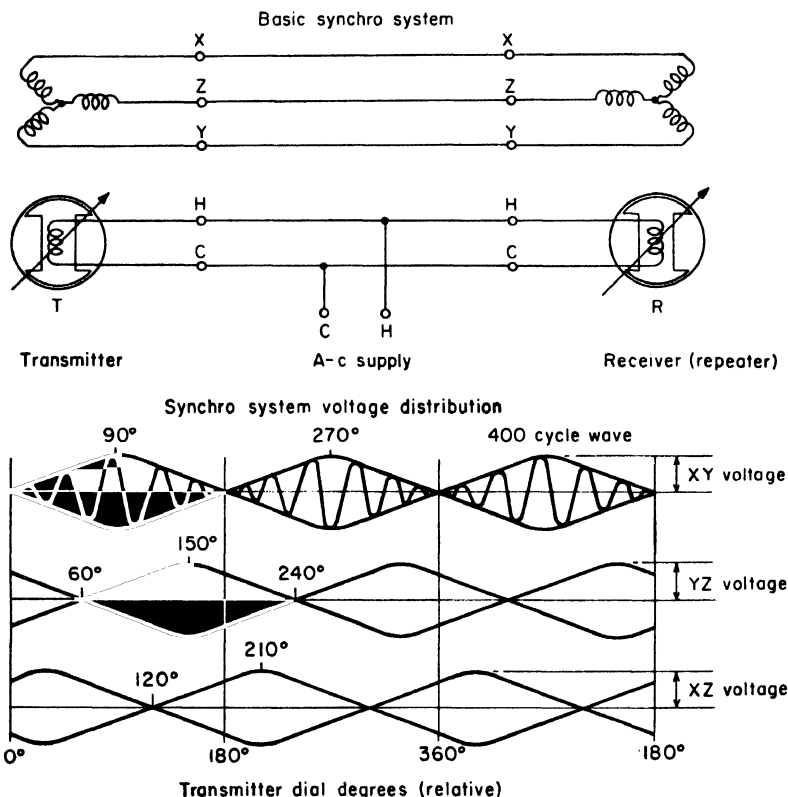



FIG. 3. A transmitter-receiver system.

A-c excitation of the synchro rotors has been standardized at 26.5 volts, 400 cps for standard synchros. The phasing of this voltage is important and the 26-volt excitation leads have been arbitrarily assigned the names "hot" (H) and "cold" (C) for the 26-volt "hot" lead and the ground lead, respectively, so that the rotor excitation may be standardized.

10. Synchro Control Transformer. A different version of the synchro system is the case in which a synchro transmitter operates with a receiver-repeater but where the receiver rotor does not have a-c power supplied (Fig. 4). Instead, the rotor is designed to supply control voltages which reach a null when the control transformer rotor is in effective alignment with the transmitter rotor. If the transmitter or the receiver rotor 's moved so as to disturb the alignment, the voltage appearing across the rotor of the ontrol transformer varies as the sine of the angle of displacement. This voltage is in ect an error signal and the voltage taken from a control transformer may be used to

actuate a servo amplifier which, in turn, drives a motor geared so as to restore the control transformer shaft to alignment. It may also be used, after further processing of the signal, to operate indicator needles which indicate left-right or other intelligence.

A servomechanism loop using an error signal from the control transformer can perform the same functions as the human pilot who responds manually to correct for changes in flight attitude. Suppose that the synchro transmitter is tied to a control surface of the aircraft and we wish to repeat the position of this control surface in an

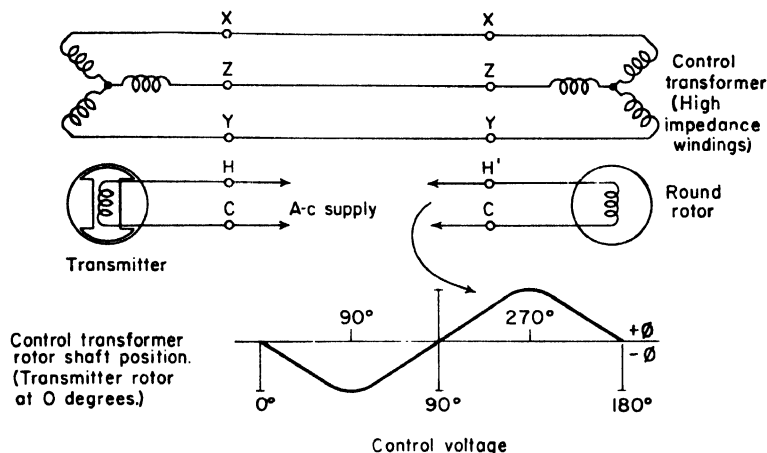



FIG. 4. Synchro control transformer.

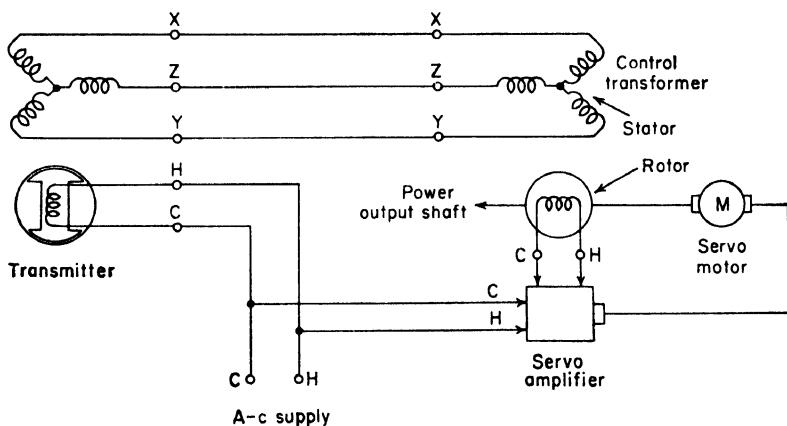



FIG. 5. Basic diagram of servosystem.

electronic device such as an automatic pilot. Assume further that turning the receiver rotor in the automatic pilot requires more power than can readily be transmitted from the synchro transmitter. In this case displacement of the transmitter shaft by the motion of a control surface results in a reorientation of the magnetic field of the synchro control transformer; this, in turn, results in a voltage being induced in its rotor. This induced rotor-error voltage is now amplified to a level sufficient to drive servomotor *M* (Fig. 5) in a direction to restore the control transformer shaft to its original alignment position. In this closed-loop system, it is apparent that the size of the

amplifier and motor may be increased virtually without limit and it is thus possible to supply large amounts of shaft power under remote control through a synchro-controlled servosystem.

11. Synchro Differential Transformer. If we wish to influence the operation of the repeater not only by the position of the transmitter synchro rotor but also by an external influence, we can introduce a *differential transformer* between the synchro transmitter and the synchro repeater (Fig. 6). In this case, rotation of the differential transformer rotor inside the stator will result in a redistribution of the X, Y, Z voltages reaching the repeater. The resultant torque on the repeater shaft and the realignment of the repeater pointer to a new angle will equal the sum or the difference (depending on the connection of the differential) of the originally transmitted angle of the transmitter rotor and the mechanical shaft angle induced through the rotation of the differential rotor. This arrangement provides a means of influencing a repeated signal from more than one information source.

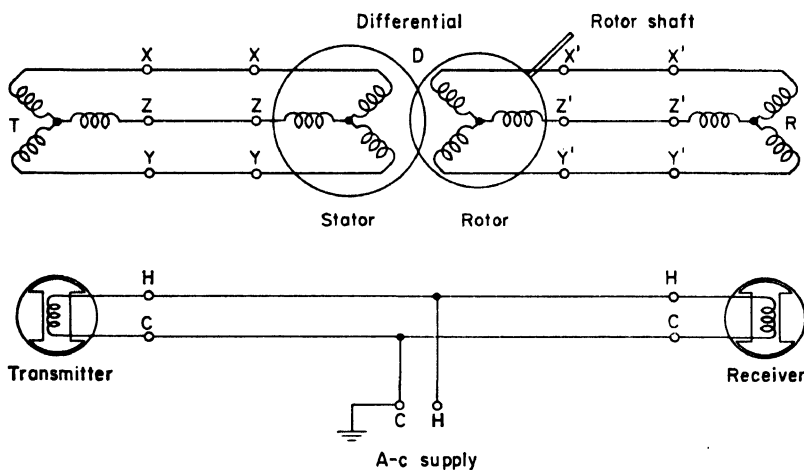



FIG. 6. Basic diagram of synchro differential system.

12. Transition to Automatic Flight. Utilizing the tools described above, we are now ready to discuss the process of automatic flight in relation to manual flight because we have described the electrical and electromechanical means to develop attitude and direction sense information required in the aircraft aloft and the methods of transmitting this information in electrical form to the automatic-flight device. Let us examine the simplest form of automatic-flight control in which we merely wish to maintain the aircraft on a straight and level course. For our reference of level flight, we utilize a vertical gyro which can render position information both in roll and in pitch from separate gimbals. By utilizing a synchro transmitter to detect the angular difference between the position of the gyro rotor wheel (always level) and the gimbal in the pitch axis (which is, in effect, attached to the aircraft) we develop a signal out of the synchro transmitter proportional to the pitch deviation of the aircraft from level flight. This deviation signal is then processed in the elevator servoamplifier and the signal is tailored so that at the output of the elevator servoamplifier is a signal which can be applied, after conversion to a mechanical force, to the aircraft elevators to correct for the pitch-attitude deviation. The signal from the servoamplifier is applied to the elevator servomotor which actuates and drives the elevator surface. A synchro transmitter located on the elevator itself reports back the position of the elevator with respect to the aircraft. If the aircraft is in level flight and the elevators are in their normal level position, the elevator synchro is at equilibrium and no error signal is sent

to the pitch-stabilization system from either the gyro or the elevators. If the elevator is not in its normal level position, an error signal is passed back to the stabilization system indicating that a pitch-attitude correction is taking place in response to an error signal generated in the gyro, thus continually reporting back response information. In this fashion, we can establish a closed-loop servomechanism between the vertical gyro pitch-axis synchro and the elevator synchro. The closed-loop system tends to maintain equilibrium with the aircraft in level flight. The vertical gyro synchro signal announcing an error in the aircraft level causes an excursion in the elevator position. This position change is reported back to the stabilization system by the elevator position synchro. At equilibrium, both synchro signals are at null and the system is at rest.

For *roll stabilization* in the aircraft, the system is very similar. In this case, the basic sensing is done by connection of the synchro to the roll axis of the vertical gyro. Now the servomotor actuates the ailerons with repeat-back synchros transmitting information on the aileron excursion back to the roll-stabilization system. Here again the roll repeat-back signal is sent back to the servoamplifier where it opposes and reduces the error signal from the roll synchro.

The *yaw reference* in modern autopilots is usually established by the signal from the magnetically slaved gyroscopic compass. This gyro spinning about a horizontal axis is similar in nature to the vertical gyro and operates in a similar fashion. A magnetically slaved gyro prevents the drift of the reference direction which would occur in a normal gyro assembly because of friction and unbalance.

When the airplane turns in azimuth, an error signal is developed in the heading synchro of the magnetically slaved gyro compass. Deviations of the airplane about the yaw axis result in rotation of the heading stator with respect to the rotor and a signal to the rudder servoamplifier, causing servoactuation of the rudder surface and the servo loop action described above.

Up to now, the automatic pilot only retains the aircraft in straight and level flight on a constant heading. By a further useful function of most *autopilots*, the pilot can select certain rates of roll resulting in controlled turns and can select rates of climb by introducing artificial error signals. Finally, the slaved gyro compass provides heading data useful as basic information to the pilot in determining his true heading as well as to the autopilot in maintaining a constant heading. Knowing his course to destination and being able to read his true heading from a heading synchro repeater located in the cockpit, the pilot can adjust the heading to make good his desired flight path.

Modern autopilots are also provided with facilities that permit navigation information to be automatically inserted to make the autopilot do additional duty. By using ground-based electronic aids, the autopilot can become the medium for automatic control of approach to an airport. Provision of altitude information from an aneroid altimeter or from a radio altimeter system will permit the autopilot to maintain the aircraft at a constant altitude.

In applying external navigation-system signals to an autopilot, the human-pilot processes remain similar to those of manual flight except that, instead of manipulating rudder pedals and the stick, the pilot may manipulate the "selector" controls of a radio receiver. The radio receiver acts as the "sensor" in detecting deviations from the ground reference chosen and produces error signals as a result of such deviations. These signals are set into the autopilot "coupler" which converts the electrical signals into mechanical energy for the "actuation" of the aircraft control surfaces which through their synchros report their position back to the autopilot. Thus the selector-control-sensor-coupler-actuator discussed earlier applies as well to the control sequence of flight by an automatic pilot as to flight by the human pilot even though the "sensors," "couplers," and "actuators" are no longer the eyes and hands of the pilot. Primarily the "selector," or decision-making element, remains in the pilot's hands.

It may be seen that the automatic pilot, while a great boon in manned aircraft, becomes an absolute requirement in the operation of guided missiles. Here, the "selector" function must also be taken over by a machine—a program computer—

which is given its directions prior to the flight and which cannot normally be corrected for changes of situation or plans after the missile has once begun its operational flight.

The internal signals from navigation aids or computers which may be used to couple additional intelligence to the automatic-flight system are discussed in subsequent parts of this chapter, as well as the technical details of specific electronic pilotage systems. In each case the basic principles of synchro systems and servomechanisms are used to couple the various signals into a useful operating mechanism.

ELECTRONIC AIDS TO AIRCRAFT MISSIONS

Electronics in aviation has gone far beyond its use as a communications medium even when communications is considered in its broadest terms to include communication, navigation, and air-traffic control. Hundreds of devices and systems have been developed to help an aircraft and its crew to perform more adequately intended missions. In the military aircraft such electronic systems as fire-control equipment, bombing systems, search radar, and electronic countermeasures are examples. In the civil-transport field engine analyzers, electronic fuel gages, fire detectors, and weather radar are used. Many of these tools are not unique to aircraft use and are merely extensions of similar use on the ground and on shipboard. Hundreds of special devices of limited applications are used in specialized types of aircraft operations.

Although it is not the intent of this chapter to delve deeply into specialized aspects of aviation electronics, a review of three or four typical devices follows.

13. The Engine Analyzer. This is a device intended to provide continuous visual analysis of an aircraft power plant. Its purpose is to detect and locate malfunctions and imminent failures that may occur during engine operation. Basically, it is an adaptation of an oscilloscope to observe the transient behavior of an engine during the combustion cycle. The voltages displayed are taken from the primary circuit of the aircraft ignition system or from a vibration pickup on the engine cylinder or any other voltage detectors depending on the engine functions to be observed. The sweep of the engine-analyzer oscilloscope is adjusted to show the vibration or firing patterns of either one cylinder or the entire group of cylinders to show a complete engine cycle. Synchronization is obtained from a three-phase generator on the auxiliary tachometer drive whose output voltage alternates in step with the engine and is used to trigger the sweep of the oscilloscope just prior to the firing of any desired cylinder.

14. Fire-control System. In a modern combat aircraft traveling at extremely high speeds, the total time during which one aircraft can be within firing range of another is extremely small. At high speeds, the time required after the completion of a pass at an enemy to get into firing position again is so long, comparatively, that a fighter aircraft may have very few chances at combat with another aircraft also traveling at high speed. Further, the time required to aim and fire a gun manually is much too long for the gunner and gun to be an efficient weapons system. For example, if a fighter aircraft traveling at 1,000 knots engages a bomber traveling at 800 knots in the opposite direction, a relative speed of 1,800 knots is attained. This means that the two aircraft are approaching each other with a relative speed of more than 3,000 fps. By the time the fighter pilot has found the enemy, taken aim, and made the decision to fire, many knots will have been traveled by both combatants and not only will the aim be incorrect but the chances are that the aircraft will be out of each other's range. To solve this problem, search radar is used to scan the sky for more miles than any human eye can see. Once an enemy is detected, the search radar locks onto him and fire-control computers go into action to compute a trajectory based on the position of the enemy's velocity and direction as well as the velocity and direction of the originating aircraft. Upon the completion of this computation the weapon may be fired automatically and without interference from the pilot. Air-borne fire-control systems are complex and critical devices utilizing not only the principles of radar but also highly refined computing devices. It is to be noted that, in the fire-control system, just as in many other electronic systems aboard the aircraft, information from the air-

craft-attitude-indicating devices (the vertical and directional gyros) is important in rendering information on the present situation of the aircraft.

15. Air-borne Weather-penetration Radar. A recent innovation on air transport and executive aircraft is the use of air-borne radar to detect and display precipitation-bearing clouds. It is the property of radar operating both in the X-band (3-cm) and C-band (6-cm) regions to display echoes not only from metallic objects or the earth but also from precipitation-bearing clouds. The weather-penetration radar takes advantage of the fact that clouds in which the precipitation gradient is high present distinctive echoes on the air-borne radar *plan-position indicator* (PPI). Clouds which have a high precipitation gradient are particularly turbulent whereas many clouds in which the precipitation gradient is low even though the total precipitation content may be high are not so characteristically turbulent. On the PPI, the various types of clouds containing precipitation present characteristic shapes so that the pilot can detect and avoid areas of heavy turbulence. By observing, on the PPI, the corridors which may exist between adjacent storm clouds, the pilot may find a calm path through what may appear to be impenetrable weather.

Peak pulse power of 50 to 75 kw is used in the weather radar, and while present-day equipment utilizes conventional cathode-ray tubes, development of bright-view storage tube indicators is nearly complete. These indicators not only have the advantage of providing greater brightness for daylight use, but also permit the displayed picture to remain on the PPI until shortly before a new sweep comes along to paint a new picture. The weather radar utilizes information from the vertical gyro in the aircraft to stabilize the antenna.

16. Electronic Countermeasures. One of the most complex and troublesome problems in military aviation is the development of electronic countermeasures and techniques to resist an enemy's electronic countermeasures. Unfortunately, most electronic systems involving electromagnetic radiation are subject to interference by other electromagnetic radiation. For example, a communication system utilizing voice as the intelligence-transmitting medium from one aircraft to another may be interfered with by a third aircraft generating noise on the same frequency on which the air-to-air conversation is attempted. Pulse systems are often susceptible to interference by continuous-wave electromagnetic radiation.

In the military sense, the development of electronic countermeasures to interfere with ("jam") an enemy's transmission and the simultaneous development of measures to prevent an enemy from interfering with our own systems involving electromagnetic radiation go hand in hand. Noise generators, frequency-seeking devices which permit rapid acquisition of the frequency the enemy might be using, and the generation of pulse groups to interfere with pulse systems are only three examples. The development of jamproof electronic radiating systems in our own aircraft is also of great importance. Electronic countermeasures as well as fire-control systems are complex and highly specialized devices. The detailed analysis of systems of this type will not be attempted here.

The above examples of only four electronic systems used in the accomplishment of aircraft function illustrate the extremely diverse nature of aircraft electronic systems.

COOPERATIVE NATURE OF COMMUNICATIONS, NAVIGATION, AND TRAFFIC CONTROL

Of the five major areas in which electronics plays a dominant role in aviation, those of communications, navigation, and air-traffic control are difficult from the systems standpoint. Whereas electronic sensing and automatic-flight equipment and other specialized electronic aids may be complex, these aids are generally independent of ground facilities and relatively independent of other aircraft. An automatic pilot is almost as useful if it is installed in only one aircraft as if every other aircraft in the vicinity is also provided with one. In the areas of communications, navigation, and air-traffic control, an aircraft is not independent of other aircraft and all are dependent on the ground facilities. Because of the many diverse classes of aircraft and ground

facilities and the need for highly coordinated activity between the ground aids and the aircraft users, the problems of providing satisfactory communications, navigation facilities, and a dependable air-traffic-control system are complex. The three categories of aids are closely allied in their application and are interdependent since successful control of air traffic from the ground requires adequate high-speed communications between the controlling agency and the aircraft to be controlled as well as precise navigation aids to assure the aircraft the ability to follow air-traffic-control instructions. This interdependence applies not only to the terminal area where large numbers of airplanes may concentrate in a relatively small geographic area prior to approach to an airport but also in the en route situation where navigation aids and electronic fixes provide the elements of safe separation and control of the traffic flow.

Given a sufficient number of ground navigation aids, an aircraft could fly from point *A* to point *B* without any communications and certainly without any air-traffic-control surveillance. Unfortunately, the aircraft could do this safely only if it were the only aircraft in the vicinity. As more and more aircraft fly in a given area, not only must navigation aids be used but communications must be provided to assure coordination of the flight paths of the different aircraft. Finally, a traffic-control agency must be provided to render the coordination service.

A whole family of electronic aids has been developed for these three functions. Some of these aids are carried aboard the aircraft while others are primarily for utilization by the ground-control agency. The provision of a satisfactory system of communications, navigation, and air-traffic control is complicated by the wide variety of aircraft types and missions to be found in any particular location. Relatively slow-speed aircraft with a minimum of electronic equipment, high-speed military aircraft in which space for this kind of electronic equipment is severely limited, and the air carriers must all share the same airspace and must, if chaos is to be prevented, utilize the same general control agency and essentially the same type of communications and navigation techniques.

The ground agency responsible for controlling the flow of such diverse traffic is faced with another problem. Electronic aids to navigation and air-traffic-control service must be provided to the major traffic hubs, such as New York, Washington, Chicago, and Los Angeles, using every possible technique available to speed communications and traffic control, and must, at the same time, provide service over relatively lightly traveled areas of the country, without compromising the basic compatibility between the types of aids used. Under the different circumstances of military missions, the problem is further complicated. While the civil communication, navigation, and air-traffic-control facilities can be used in peacetime and within the established ground-facility arrangements all over the world, in wartime the provision of this service becomes complicated because the traffic hubs are mobile and navigation aids as well as communication and traffic-control facilities are subject to rapid and violent changes of location and emphasis.

Subsequent sections of this chapter will deal with the areas of communications, navigation, and traffic control separately, but the reader will do well to keep in mind the interdependence of the three types of facilities.

THE DEVELOPMENT OF STANDARDS FOR AVIATION ELECTRONICS FACILITIES

The several groups of users of the airspace each have individual requirements for electronic aids to communications, navigation, and air-traffic control. If each of the individual user groups were to be supplied with specialized tools to do its own job, our system of airways would be chaos. The airways must be used in common by air-line flights, nonscheduled operators, personal fliers, business fliers, and military forces.

In 1939, there were 27,074 miles of airways implemented with facilities for radio navigation and communications. Control towers were in operation at 52 locations to control the aircraft on and into the vicinity of airports, and 11 air-route traffic-control centers were in operation to control aircraft conducting flights on the airways. The

Second World War was a great stimulant to progress. New equipment was developed and rushed into action. The need was not only for tactical equipment for the fighting zones, but also for improvement of the airways in the United States since every type of aircraft had to move on these airways under poor weather conditions and with all possible speed.

17. The "Common System." The expanding economy of aviation made it quite obvious after the Second World War that the different user groups could not be individually served by different systems of aids, and it was agreed that the various interested private groups and responsible government agencies had to reach agreement on a "common system" of communications, navigation, and traffic control which would adequately serve both civil and military needs. The President-appointed Air Coordinating Committee, organized to investigate aviation problems and developments affecting more than one government agency, agreed that a single group should attempt to resolve the airway problems. The job was undertaken by the Radio Technical Commission for Aeronautics, a cooperative association of United States government and industry aeronautical telecommunications agencies, both military and civil. RTCA, through its Special Committee 31, laid down basic specifications for a common system and proposed a program and equipment to meet them.

The basic principles on which the common system was to be based were as follows:

1. *Safety.* The operation of all aircraft requires that adequate separation of aircraft be maintained at all times and under all conditions.
2. *Expeditiousness.* The expeditious movement of aircraft requires airspace utilization under all conditions equal to that under VFR[1] conditions.
3. *Reliability.* The system shall, as far as possible, be independent of weather conditions and equipment breakdown.
4. *All Aircraft Use.* The system shall be usable by all types of aircraft in a manner compatible with their performance characteristics and operational applications.
5. *All Weather.* The system shall function effectively during all weather conditions.
6. *All Airspace.* The system shall be capable of expansion to include all usable airspace.
7. *Minimum Control.* The system shall provide control that imposes a minimum of restriction on the movement of air traffic.
8. *Integration.* The elements of this system shall be integrated.
9. *Human Factors.* The system shall impose a minimum burden on such human factors as competency, training, and alertness of air and ground personnel.
10. *Evolution.* The system shall permit orderly transition from the utilization of established air-traffic-control aids to those ultimately required.
11. *Flexibility.* The system shall incorporate characteristics which make practicable the handling of a wide variety of operational situations.
12. *Security.* The system shall be designed for due consideration for system security in the interest of national defense.
13. *Limitation of Traffic Flow.* The system in itself shall not limit traffic flow below the acceptance rate of the airports.
14. *Language Difficulties.* The system shall operate independently of the language capabilities of ground and airborne personnel.
15. *Identification.* The system shall provide a means of identifying aircraft as required for purposes of traffic control and in the interest of national security.
16. *Division of Responsibility.* The Traffic Control Agency shall be responsible for air-traffic-control planning, the collection of pertinent information, and the dissemination of traffic clearances and information.

   The pilot shall be responsible for compliance with the control agency's clearances, subject to a primary responsibility for the safe conduct of the flight.

   The appropriate regulatory authority shall promulgate rules and regulations required for the proper functioning of the air-traffic-control-system.
17. *Division of Equipment.* The system shall place the primary burden of weight, volume, and physical complexity on the ground. The aircraft equipment shall result in the minimum detriment to aircraft performance.
18. *Frequency.* The frequency allocation for the system shall be consistent with the National and International Tables of Frequency Allocations.
19. *Cost.* The cost of the system shall be compatible with the results obtained.

[1] "Visual flight rules."

Special Committee 31 was extremely successful in winning acceptance of its findings from all the previously divergent groups, including military aviation, the air lines, the nonscheduled operators, the executive operators, and the private fliers.

Special Committee 31 of RTCA did its work in 1947 and 1948, and while the principles laid out are as valid today as they were then, implementation of the elements of the common system has been a slow process. While there has been general agreement on the job to be done and the principles involved, controversies concerning the implementation of specific tools within the common-system framework have occurred between the various users with divergent interests. In spite of delays, however, progress has been made.

In resolving the problems of the common system, a number of government and nongovernment agencies have a direct stake. The responsible agency for the nation's Federal Airways System, the comprehensive network of navigational facilities, communications, traffic control, and the flight advisory services is the Civil Aeronautics Administration. The Congress has authorized and directed the Administrator of Civil Aeronautics "to designate such civil airways as may be required in the public interest and authorizes him to develop, establish and improve, operate and maintain air navigation facilities wherever necessary, to provide the facilities and personnel for the regulation and protection of air traffic and in the exercise of his authority to give full consideration to the requirements of National Defense." The operation of the Federal airways is an essential element of both military and civil aviation. The common system, however, cannot be the responsibility of any one agency alone. Since it must serve all military nontactical aviation and international aviation as well, a number of different agencies in addition to CAA have a direct stake and responsibility in the development of the common system. Common-system developments are characterized by the tremendous amount of coordination required for their successful implementation, and literally dozens of government agencies, government-industry groups, and industry groups are directly concerned with its development and implementation.

18. The Responsible Agencies. While a separate volume would be required to describe the many agencies involved in the development of the common system and a small library to describe the coordination processes, a few of the most important national and international government agencies who have responsibility in the systems of communications, navigation, and air-traffic control are described.

The five-member Civil Aeronautics Board (CAB) exercises the functions of rule making (including the prescription of rules, regulations, and standards), adjudication, and investigation concerning civil aeronautics. The Administrator of Civil Aeronautics and the CAB constitute the Civil Aeronautics Authority. However, the Civil Aeronautics Authority, as such, performs no functions.

19. The Civil Aeronautics Administration. "The Administrator of Civil Aeronautics encourages and fosters the development of civil aeronautics and air commerce; encourages the development of civil airways, landing areas, and other navigation aids and facilities; designates civil airways and acquires, establishes, operates, and maintains, air navigation facilities along such civil airways and landing areas; makes provision for the control and protection of air traffic moving in air commerce." The activities in air navigation cover the planning, construction, maintenance, and operation of the Federal Airways Systems. The duties of the CAA include the purchase and installation of radio and communication equipment, operation of the air-traffic-control system, maintenance and operation of government-owned air-navigation facilities on the Federal Airways System, conduct of research for development and improvement of new types of equipment to air-navigation aids, aircraft power plants, flight instruments, etc., and furnishing assistance to the Departments of Army, Navy, and Air Force, including surveys, selection of sites, construction, and installation of radio range and communication facilities at points throughout the domestic and foreign zones of military operations.

20. The Air Coordinating Committee. This committee consists of members from the Department of Commerce, the Civil Aeronautics Board, the Department of State, the Department of the Army, the Department of the Navy, Department of the Air Force, Department of the Treasury, Department of the Postmaster General, the Federal Com-

munications Commission, the Bureau of the Budget, and the Office of Defense Mobilization. The committee (ACC) considers joint problems affecting more than one participating agency. ACC develops and recommends "integrated policies to be carried out and action to be taken by the participating agencies or by any other government agency charged with the responsibility in the aviation field and to the extent permitted by law, coordinates the aviation activities of such agencies except activities relating to the exercise of quasi-judicial functions." ACC provides liaison between government and industry activities in the aviation field and also consults with representatives of the United States to the International Civil Aviation Organization and with Federal interagency boards and committees concerned with aviation activities.

21. The International Civil Aviation Organization (ICAO). ICAO is a "specialized agency of the United Nations. The purposes of ICAO are to develop principles and techniques of international air navigation, to foster the planning and development of international air transport so as to insure the safe and orderly growth of international civil aviation, encourage the design and operation of planes for peaceful purposes, guard against economic injustices in international air transport, promote safety and in other ways promote the development of international civil aviation."

"Through ICAO the member states seek to achieve uniformity in all matters where uniformity will facilitate and improve air navigation. Uniformity is achieved through the adoption of Annexes to the Convention known as 'Standards and Recommended Practices.' These Annexes deal with such matters as air navigation aids, airports, rule of the air, licensing of personnel and meteorological information."

22. The Federal Communications Commission. The FCC was created for the purpose of "regulating inter-state and foreign commerce in communications by wire and radio so as to make available as far as possible to all people of the United States a rapid, efficient nation-wide and world-wide wire and radio communication service with adequate facilities at reasonable charges, for the purpose of the national defense, for the purpose of promoting safety of life and property through the use of wire and radio communications."

In aviation, where communication and navigational facilities employing electromagnetic radiations are widely used and vital, the FCC Aviation Division is concerned with the efficient utilization of radio frequencies and the radio spectrum by all classes of aviation. By virtue of this responsibility the FCC is vitally interested in the development and regulation of systems of communications, navigation, and traffic control.

23. The Airways Modernization Board. In 1957, Congress established the Airways Modernization Board with an assigned responsibility to develop, test, modify, and evaluate systems, procedures, facilities, and devices, as well as to define the performance characteristics thereof. It is empowered to select such systems, procedures, etc., as will best serve such needs and will promote maximum coordination of air-traffic control and the air defense system.

The Board is composed of voting members from the Defense and Commerce Departments and a President-appointed chairman. It is envisioned that the AMB, an interim organization, will become the research and development arm of a Federal Aviation Agency, which would consolidate all the essential management functions necessary to support the common needs of military and civil aviation of the United States, except certain air warfare functions.

24. Radio Technical Commission for Aeronautics. RTCA was organized in 1935 upon invitation of the Assistant Director of Air Commerce of the Department of Commerce. Government and industry agencies concerned with the development, application, and use of radio in aeronautical operations were invited to form an organization for the coordination of effort in the technical development of the uses of radio in aeronautics.

RTCA is a cooperative nonprofit association of the United States government-industry aeronautical telecommunications agencies. It conducts studies of aeronautical telecommunications problems and related matters. Its objective is the resolution of such problems by mutual agreement of its member agencies. Its findings are in the nature of recommendations to all United States organizations concerned. The RTCA is not an official agency of the government. Its recommendations, therefore, may not be regarded as statements of official government policy unless so enunciated by the government agencies having statutory jurisdiction in the matters to which the recommendations relate.

25. The International Telecommunication Union. ITU is a specialized agency of the United Nations, having a membership of 90 countries and 5 associate members. The association originated from the International Telegraph Union established in 1865 and resulted from a merger in 1932 into one convention of provisions pertaining to radio, telegraph, and telephone; the 1932 convention was revised at the Atlantic City Conference

in 1947 and at the Buenos Aires Conference in 1952. "The purpose of the ITU is to establish the most efficient and economical systems of telecommunications possible throughout the world, through regulations governing the international use of telegraph, telephone and radio services and through technical and scientific studies designed to improve the means of communication." In the radio field, the chief activity of the ITU has been an attempt to work out, through cooperative procedures, an orderly and equitable use of frequencies in the limited radio spectrum. Periodic international conferences, under ITU sponsorship, have been held to consider regulatory matters and frequencies involved in international fixed (point-to-point) radio, aeronautical operations, high-frequency broadcasting, maritime-service communications, etc.

26. Joint Chiefs of Staff–Joint Communications Electronics Committee. The Joint Chiefs of Staff prepare strategic plans and provide for the strategic direction of the military forces, prepare logistic plans, and assign to the military services logistic responsibilities in accordance with such plans. With the tremendous importance of electronics to the military services, the Joint Communications Electronics Committee advises and formulates plans and policy as they apply to communications and electronics for both military aviation and other services.

27. USAF Directorate of Communications-Electronics. In the Air Force, the Deputy Chief of Staff, Operations, is responsible among other things for Air Force communications activities. The Directorate of Communications-Electronics is directly concerned with aviation-electronics matters.

28. The Air Research and Development Command. ARDC is responsible for the discovery, evaluation, development, and testing of qualitatively superior means for the accomplishment of the Air Force mission, including that of aviation electronics.

29. Air Defense Command. The Air Defense Command is charged with providing for the air defense of the United States and is vitally concerned in aviation matters particularly as applies to the identification, by electronic and other means, of aircraft operating within the United States and particularly near the point of entry.

30. Office of Defense Mobilization. Within the organization of the Office of Defense Mobilization are two coordinating groups having responsibilities which encompass aviation-electronics activities—namely, the Inter-department Radio Advisory Committee and the Telecommunications Planning Advisory Committee. The membership of these groups includes senior communications officials of departments and agencies of the government. They are charged with the development and coordination of telecommunications plans, programs, and actions under work programs established and adopted by the Telecommunications Planning Committee. An important phase of the duties of this organization is to coordinate the assignment and use of radio frequencies, particularly within the branches of the military service.

These are only a few of the many agencies and groups having a vital interest in aviation electronics. It is enough of a cross section to indicate the complexity of developing coordinated plans and programs suited to all these groups. In addition to specific responsibilities for its own phase operation in aviation, each of them must coordinate its activities within the framework of the common system.

AVIATION COMMUNICATIONS

The most venerable use of electronics in aviation is communication by radio. Communication forms the backbone on which all coordinated movement of air traffic depends. It is the link through which aircraft receive information on weather, air-traffic control, clearances, and all other messages bearing on the operation of particular services. Thus communication services are a major part of the system of Federal airways. Further, extensive aviation-communications systems are maintained by the military services for strategic and tactical purposes and by the air-transport companies for operational control of their aircraft.

It is readily apparent that, for the control of air traffic, radio contact between the aircraft and a number of ground facilities is not enough. Efficient operation of air traffic depends both upon air-ground-air communications and upon a rapid means of exchanging intelligence between air-traffic controllers, between the controllers and personnel in flight, dispatch offices, flight services, service centers, communication stations, search and rescue, and air-defense facilities.

The CAA system of Federal airways operates a number of different communication services, both point-to-point (fixed service between ground locations) and ground-air/air-ground facilities.

31. Point-to-point Communications Services. Communications acceptable for transmission and the preferred system for each type of communication are described below:

Service A

1. Emergency communications (of any agency engaged in aeronautical activities) which pertain to or affect the safety of life or property.
2. Meteorological information.
3. Notices to airmen.
4. Meteorological information ordinarily routed on services C and O if offered during periods of communication failure.
5. Aircraft movement and control messages, when existing systems provided for this purpose have failed.
6. Communications having a direct bearing on the maintenance or operation of air-navigation facilities, including official administrative messages of the CAA. This type of communication is authorized only when use of no other communication service, either governmental or commercial, would suffice.

This service is composed, at present, of 12 teletypewriter circuits, each covering an appropriate portion of the United States and having from 100 to 300 terminals (drops) of which about 50 are send-receive and the remainder receive only (total drops, nationwide, about 1,800). The operational weather broadcast on this system consists of hourly service reports, special reports, radar reports, pilot reports, supplementary reports, and miscellaneous data, including a limited amount of upper-air and forecast information required for pilots and dispatchers for operational purposes. Notice-to-airmen information consists of information and advice regarding the status of radio aids to air navigation, airports, and airway lights.

Service B

1. See item 1, service A.
2. Aircraft movement and control messages.
3. Notices to airmen, provided service A is not operative, or when delays, if held for service A transmission, would be excessive.
4. Meteorological information during the failure of other facilities.
5. Communications having a direct bearing on the operation of air-navigation facilities, including official administrative messages of the CAA. This is the preferred type of system for this type of communications.
6. Certain messages filed by Federal law-enforcement agencies.

Teletype circuits are used also to meet the requirements for exchanging intelligence between controllers and between controllers and personnel in flight-dispatch offices, etc., but only to a limited degree for the transmission of control messages, because of current relatively slow operating speeds. It is also used for transmission of international flight safety messages between OFACS. The service B network provides a circuit to serve each air-traffic-control center's area and to provide certain essential "express" circuits and automatic relay between all circuits. The service B speed of operation is in the process of being increased from 75 to 100 wpm. Extensive use is expected to be made of service B for the exchange of less critical control messages between the air-route traffic-control centers and thus relieve congestion and permit more rapid exchange of the time-critical messages on service F.

Service C

1. See item 1, service A.
2. Meteorological information.
3. Meteorological information ordinarily routed via services A and O if offered during periods of communication failure.

4. See item 5, service A.
5. See item 6, service A.

Service C is composed of six teletype circuits, each covering a selected area of the United States and averaging about 75 drops in the offices of a widely diversified group of users.

In the main the data consist of synoptic information used for forecasting and general public weather service. (This service includes periodic synoptic reports, winds-aloft data, terminal, regional, and state forecasts, and various types of weather analyses.)

Service F

1. See item 1, service A.
2. Communications essential for the efficient control of air traffic.
3. Aircraft-movement information offered during periods of failure of service B. Individual weather reports ordinarily routed via other systems, if offered during periods of communications failure.
4. See item 6, service A.

Service F using interphone is primarily employed for time-critical messages. Some of the requirements for exchange of intelligence between controllers is met by talking directly, by telephone, between operating positions within a center or tower. The interphone service F messages have made it possible to segregate control messages from routine aircraft-movement messages, the interphone being used for control messages and the teletypewriter for the transmission of aircraft-movement messages between centers.

Service O

1. See item 1, service A.
2. Meteorological information.
3. Meteorological information ordinarily routed via services A and C if offered during periods of communications failure.
4. See item 5, service A.
5. See item 6, service A.

Service O is composed of 5 circuits serving 90 locations. The data transmitted over service O consist of operational and synoptic weather data received from or having to do with locations beyond the continental limits of the United States which require delivery to key international terminals and large forecast centers within the United States.

32. Ground-Air-Ground Communications Facilities. CAA also renders several different services air to ground and ground to air:

The Air Traffic Communications Station (ATCS). An ATCS is a specialized aeronautical communication station operated by CAA to provide communication between pilots and air-route traffic-control services and to provide flight-information and flight-assistance services primarily to military, itinerant, and other non-air-carrier users. Other functions of the ATCS's are the operation of a land-line system for the collection and dissemination of weather information and the exchange of aeronautical safety messages and to monitor the performances of nearby navigational aids and to initiate remedial actions in case of their malfunctioning.

The development of the ATCSs began with the establishment of low- to medium-frequency range facilities to provide voice weather broadcasts over the radio-range frequency. This soon expanded to two-way communication using the voice channel of the radio range for the ground-to-air channel. As activity increased, separate air-ground communications, independent of the navigational service, were added. Coordination between such stations was provided by secondary use of the point-to-point teletypewriter system which, in turn, was later supplemented with separate services for each of these functions.

The ATCSs became an indispensable part of the air-traffic-control system, particularly as contact points for military and non-air-line civil aircraft. Some of the activities of ATCSs may be displaced with the increase in direct center-to-aircraft communications.

As an indication of service rendered by interstate airway communication stations, 6,817,657 individual aircraft contacts were made during fiscal year 1955. Of these con-

tacts, nearly 4 million were aircraft operating under instrument flight rules (IFR) and defense visual flight rules (DVFR). The distribution of these aircraft was 71 per cent military, 11 per cent air carrier, and 18 per cent other civil. The small percentage of air-carrier aircraft contacted through ATCSs results from the fact that the air-line needs are met by the nongovernment air-transport-company facilities.

International Air Traffic Communications Station (IATCS). IATCS is a CAA aeronautical telecommunications facility for maintaining aeronautical communication services between the United States and its territories and possessions and foreign locations in the major world-air-route system. These stations provide the means for collection, dissemination, and exchange of international meteorological data, the handling of the necessary categories of aeronautical safety communications between fixed points on international air routes, and the handling of air-ground communications associated with international air operations. Ten major facilities are now in operation by the CAA, and these facilities provide services in accordance with standards established by the ICAO. These facilities fulfill a part of the United States responsibilities under ICAO international agreements as a portion of the world-wide service to aviation coordinated by that organization.

33. Direct Center-to-pilot Air-to-ground Communications. This facility provides for the use of air-ground communication channels by air-traffic controllers in air-route traffic-control centers for direct contact with pilots for the purpose of expediting the control of en route traffic. In the control of traffic, particularly in congested areas where there may be a large number of high-speed and low-speed aircraft intermixed, the use of direct communications between controller and the pilot has proved to be an efficient way of conveying traffic-control instructions in minimum time. Direct communications permit more rapid pilot acknowledgment and compliance with ATC instructions than is possible when traffic-control instructions are relayed between controller and pilot via an intermediary, and permit the handling of a greater number of aircraft. The direct controller-pilot communication system also expedites the control of air traffic in that it permits direct and rapid coordination between the two parties required in the solution of special or complex control problems. This program is an extension of the direct controller-to-pilot communication system originally inaugurated in airport traffic-control towers to expedite the movement of traffic in the terminal area. It is planned ultimately to provide each control position in the air-route traffic-control centers with a discrete radio-communication outlet to the aircraft under its control responsibility.

34. Operational Control Communications. Several major segments of aviation maintain extensive networks of communications for operational control for the safe and expeditious conduct of the airway users' flights and for the essential communications required to conduct their business.

The U.S. Air Force Communication System.[1] One example of such a system is that in use by the U.S. Air Force, the U.S. Strategic Communications System. This system, separate from the CAA system of Federal airways, encircles the globe with radio stations, land lines, switching centers, and ground-air stations. The domestic United States portion of the system consists of "five major semi-automatic teletype switching centers connected by leased wire trunk circuits. Each switching center has its satellite air base tributary stations. In addition, the USAF Strategic Air Command (SAC) has special leased telephone, facsimile, and teletype lines directly interconnecting United States bases. Important leased-wire lines are backed up by high frequency radio facilities."

The overseas communications lines through the Atlantic, Mediterranean, Middle East, the East, and the Pacific are high-frequency four-channel time-division-multiplex systems. Conversion of some of these systems to single-side-band plus multichannel teletype has been undertaken. The circuits through the North Atlantic area to Labrador, Greenland, and Iceland are v-h-f beyond-the-horizon systems which have proved more reliable than high-frequency circuits. All the v-h-f circuits are either four- or eight-channel multiplex but newly developed equipment may make it possible to double this channel capacity.

The Air Force global communication system handles over 3½ million teletype messages monthly, plus some quarter million ground-air contacts with military aircraft.

[1] Donkin, F. W., Col., USAF, "U.S. Air Force Communications Systems Problems," IRE 1956 National Convention.

Civil Air Transport Company Communications. Part 40 of the Civil Air Regulations concerning communications facilities requires that "the air carrier shall show that a two-way air-ground radio communication system is available at such points as will ensure reliable rapid communications over the entire route either direct or via approved point-to-point circuits for the following purposes:

"(*a*) Communications between airplanes and the appropriate dispatch office, in which case such systems shall be independent of systems operated by the Federal Government, and

"(*b*) Communications between airplanes and the appropriate air traffic control unit, in which case the Administrator may permit the use of communications systems operated by the Federal Government."

To meet the requirements for a separate communications system for company operational control, more than 500 nongovernment air-ground-air communications stations and associated point-to-point land-line facilities are used by air-line companies. This system provides continuous contact between the offices of each air-line company and their aircraft in flight and is utilized for the transmission of CAA flight instructions where these are not covered by direct CAA aircraft communications, for the exchange of weather-advisory information, and the normal details of transport operation. The transport companies also operate teletype circuits and point-to-point radio stations interconnecting their administrative offices for the exchange of communications essential to the safe, expeditious, and economical operation of high-speed mail, cargo, and passenger service.

Aeronautical Radio, Inc. The communication activities of the civil air transport companies are coordinated through the facilities of Aeronautical Radio, Inc. (ARINC), a corporation in which the United States scheduled air lines are the principal stockholders. Aeronautical Radio, Inc., is the FCC licensee of the nongovernment aeronautical ground radio stations. The majority of these stations are located on air-line property and are operated under lease-contract arrangements by the using air lines themselves. A number of these stations, mainly at major points of entry into the United States, are operated by ARINC personnel. All of them serve any air line or other aircraft having contractual arrangements with ARINC for the operation, maintenance, and shared liability of the respective facilities.

ARINC is also the coordinating agency for air-line-industry activities in communications matters. In addition to air-ground-air service from more than 500 locations throughout the United States, ARINC furnishes the international air carriers and others, including the United States government, point-to-point communications service to overseas and foreign points through ARINC-operated aeronautical fixed stations.

The operation of the ARINC air-ground-air service renders to the air lines, and to other contract users, a service which is used in most instances in lieu of the CAA ATCSs. It maintains continuous contact between the aircraft and the appropriate dispatch office for messages of in-flight advisory nature and operational control, although a large part of such communications relate to air-traffic control.

35. Other Aeronautical Services. Several aeronautical services not described in the preceding sections are of interest:

Airdrome Control Stations. Communications of an airdrome control station are intended for the safe and expeditious operation of aircraft using specific airport facilities for operating within an airport control area. The services of the airport control stations include those at the control tower at air-route traffic-control centers for communications with ground vehicles and aircraft on the ground at airports.

Aeronautical Utility Mobile Stations. Using some of the frequencies available to airdrome control stations, communications by an aeronautical utility mobile station takes care of the necessities of ground-traffic control at an airport and is used for essential communications with control towers, ground vehicles, and aircraft on the ground.

Aeronautical Public-service Stations. Certain frequencies available to ship telegraph and ship telephone stations are available also to aeronautical public-service aircraft stations for the handling of public correspondence. Although an adequate public telephone service is not available on a regular basis to the scheduled air carriers or other aircraft operators, planning fro such service as an adjunct to the public long-distance telephone service is under way.

Civil Air Patrol Stations. A number of frequencies are available for the routine and emergency needs of the Civil Air Patrol.

Aeronautical Advisory Stations. These stations may be provided by the owners or operators of landing areas not served by airdrome control stations, for the purpose of communicating with aircraft to advise on the necessities of safety and expeditious operation of aircraft. These stations handle communications primarily concerned with the condition of runways, types of fuel available, wind conditions, weather information, or other information necessary for aircraft operation.

Aeronautical Metropolitan Stations. These stations are intended to provide communications on a limited-area basis for the safe, expeditious, and economical operation of aircraft operating between the main air terminal of a metropolitan area and subordinate landing areas. Such stations normally communicate primarily with aircraft but are authorized secondarily to intercommunicate with other aeronautical metropolitan stations within the same metropolitan area. This service is particularly suited to the coordination of helicopter operations in a metropolitan area.

36. Military Air-traffic-control Communications. While the majority of civil air-traffic control and domestic en route v-h-f communications are confined to the band 108 to 127 Mc (108 to 118 Mc is reserved for radio navigation stations, the v-h-f Omnirange and the Instrument Landing System Localizer Service), military communications are carried on predominantly in the band from 200 to 400 Mc. This was planned to provide one continuous frequency spectrum for interchange of communications among various military services. A large number of channels for this purpose, as well as for the normal military air-traffic control, are available in this band. The transition by the military to the u-h-f spectrum is proceeding rapidly. The CAA, which has the responsibility for communications and air-traffic-control services to military as well as civil aircraft flying the Federal airways, provides a companion u-h-f channel for each v-h-f channel so that, in effect, the v-h-f and u-h-f channels assigned to a particular station function as one station. This provides the advantage of eliminating a conflict between different controllers in use of v-h-f and u-h-f channels, and where this pairing is used, all aircraft in a particular control area receive information concerning other traffic in the area regardless of whether the aircraft are equipped with v-h-f or u-h-f equipment.

FREQUENCY ASSIGNMENTS IN THE AERONAUTICAL SERVICE[1]

37. Airdrome Control Stations. A number of frequencies in the band 118.1 to 126.3 Mc are assigned to this service. These assignments, originally made with a channel spacing of 200 kc on odd-tenth megacycle channels, are now also being assigned on 100-kc spacings as more selective air-borne equipment is becoming available.

121.7 and 127.9 Mc are available for communications with ground vehicles.

121.5 Mc is designated as the universal channel for emergency and distress communications.

278 kc is available for assignment in addition to the v-h-f service.

38. Aeronautical En Route Stations. These stations provide all necessary non-public-service h-f and v-h-f service to any aircraft-station licensee who makes cooperative arrangements. In case of emergencies, these stations provide the above service without prior arrangements.

Several families of frequencies in the band 2 to 20 Mc are assigned for domestic routes in the continental United States.

Groups of frequencies, deployed on a regional basis in the continental United States, are assigned to v-h-f chains in the band 126.9 to 132 Mc.

Other families of frequencies in the h-f band are assigned to the international high-frequency service serving the major world-route areas.

The v-h-f frequency 126.9 Mc is available to aeronautical en route stations serving international operations.

[1] For detailed breakdown of actual frequencies assigned, see Rules and Regulations of the Federal Communications Commission, Part 9, Aviation Services, effective Dec. 14, 1956. Because of the extreme mobility of modern aircraft and the need for flexibility in aeronautical radio operations, the assignment and utilization of frequencies must be coordinated on an international basis. This is accomplished through the ICAO and the ITU (see Secs. 21, 25). National regulations, such as the FCC rules, reflect such international agreements in the use of frequencies in the aviation service.

39. Operational Fixed Stations. These stations are authorized for link or control service or other aeronautical fixed operations.

They may be assigned frequencies in the bands 72.02 to 74.58 Mc at each 0.04 Mc, such as 72.02 Mc, 72.06, etc., and in the band from 75.42 to 75.98 Mc at each 0.04 Mc, such as 75.42, 75.46, 75.50 Mc, etc.; 75 Mc is designated for aeronautical marker beacons.

Operational fixed stations may also be assigned frequencies in the following bands:

952 to 960 Mc
1,850 to 1,990 Mc
2,110 to 2,200 Mc
2,500 to 2,700 Mc
6,575 to 6,875 Mc
12,200 to 12,700 Mc

40. Aeronautical Utility Mobile Stations. Very high frequencies of 121.7 and 121.9 Mc are available.

41. Radio-navigation Stations. *Localizer Stations.* 108.1- to 111.9-Mc channels, separated by 200 kc, such as 108.1, 108.3, and 108.5 Mc.

Glide-path Stations. The band 328.6 to 335.4 Mc.

Aeronautical-marker-beacon Stations. 75 Mc.

Omnidirectional Radio-range Stations. 112.1 through 117.9 Mc and the following frequencies in the 108- to 112-Mc band: 108.2 Mc to 112.0 Mc spaced 200 kc.

Radio-beacon Stations. 200 to 400 kc.

NOTE: Air-navigation facilities are usually operated by the CAA, although the frequencies which these facilities employ are available for licensing by the FCC upon a showing of need where the government is not prepared to render such service.

42. Flight-test Stations. V-h-f and h-f frequencies are available for flight-test stations shared with flying-school stations on a noninterference basis. A number of frequencies in the 200- to 400-Mc band are available for flight-test-station telemetering activities.

43. Flying-school Stations. 123.1, 123.3, 123.5 Mc shared with flight-test stations on a noninterference basis.

44. Aeronautical Public-service Stations. Frequencies available to ship telegraphic and ship telephone stations are available to aeronautical public-service aircraft stations for the handling of public correspondence.

45. Civil Air Patrol Stations. Several v-h-f and h-f frequencies are available for land and mobile stations of the Civil Air Patrol.

46. Aeronautical Advisory Stations. 122.8 Mc is available for communications with private aircraft stations and for communications with private aircraft engaged in organized civil-defense activities.

47. Aeronautical Metropolitan Stations. Frequencies available for aeronautical en route stations are available for assignment to aeronautical metropolitan stations.

NOTES: *Frequency-stability Requirements.* The FCC rules require that the carrier frequency of stations in the aeronautical services shall be maintained within the following percentage of the assigned frequency (this does not apply to certain devices such as radio altimeters, radar, etc.):

(*a*) All aircraft stations on frequencies above 500 kc	0.01
(*b*) All ground stations on frequencies above 30,000 kc	0.01
(*c*) All stations on frequencies of 500 kc or below	0.02
(*d*) Aeronautical fixed stations on frequencies from 1,605 to 4,000 kc	
With power of 200 watts and below	0.01
With power above 200 watts	0.005
(*e*) Aeronautical fixed stations on frequencies from 4,000 to 30,000 kc	
With power of 500 watts and below and with power above 500 watts	0.003
(*f*) All other ground stations on frequencies from 1,605 to 4,000 kc	
With power of 200 watts and below	0.01
With power above 200 watts	0.005
(*g*) All other ground stations on frequencies from 4,000 to 30,000 kc	
With power of 500 watts and below	0.01
With power above 500 watts	0.005

In general practice, much tighter tolerances are maintained.

H-f Frequency Assignments. Because of the great variations in high-frequency propagation conditions at various times of the year, under different conditions of sunspot activity, and under many short-term propagation conditions, various families of frequencies in the

h-f band are assigned to the several services to permit flexibility and continuous service under the varying conditions. A typical family of frequencies would consist of three or more frequencies from the 3-, 5-, 8-, 11-, 13-, and 17-Mc orders.

V-h-f Channel Spacing. Although the present channel spacing is 200 kc in the band 126.8 to 132.0 Mc, additional intermediate frequencies at 100- or 50-kc spacing may be authorized for use if it is shown that harmful interference will not be caused to aeronautical en route stations operating on v-h-f frequencies. Frequencies in the air-traffic-control service in the 118.1- to 126.3-Mc band, while normally assigned with 200-kc spacing, may also be deployed on a 100-kc spacing basis according to criteria established by a special committee of RTCA. These more closely spaced channels as well as the 50-kc spaced channels falling between them will be used with due regard to interference protection as more selective air-borne receiving equipment becomes generally available.

48. Recent Developments and Trends in Aeronautical Communications. Both the h-f spectrum and the v-h-f spectrum are extensively used. While the h-f service is extremely important, particularly for long-haul operations where other methods of communications are not feasible or satisfactory, the v-h-f service offers better intelligibility, freedom from static, less exposure to interference, and more positive geographical coverage even though confined to line of sight. With the domestic United States, the program for the CAA air-traffic-control-system expansion anticipates the continued and expanded use of v.h.f. for civil operations and the u-h-f spectrum for military operations. With the limited amount of frequency spectrum available for civil v-h-f communications, the problem of obtaining additional frequencies to serve the ever-expanding aviation industry becomes a serious one. A long-range plan, coordinated several years ago by RTCA, which had to await technical and economic capability, is now being implemented. This provides more u-h-f frequencies by channel splitting, thus utilizing the intermediate frequencies spaced 100 kc instead of the 200-kc spacing previously used. A further step, to 50-kc spacing, is contemplated as economics permits. Both CAA and the air carriers are taking these steps to afford useful increase in the total number of available channels and the more efficient utilization of that portion of the spectrum available to aviation.

49. V-h-f Networks. Voice communications by v.h.f. is naturally limited by the line-of-sight character of v-h-f radiation. Where extensive en route coverage is required of v-h-f communication service, careful siting of ground facilities is employed to assure the desired route area coverage. Ground stations may be used singly or in tandem. One method involves the simultaneous transmission and reception from the various ground stations, so that aircraft may be assured at any time of receiving signals from one or more ground transmitters, each of which carries the same message content. Such interconnected chains of stations are in use in Europe for air-traffic-control networks and are being implemented in the United States for certain segments of the air-transport industry.

The same nominal v-h-f frequency is used for several interconnected ground-network transmitters which transmit the same information simultaneously. The discrete frequencies used by the several transmitters are offset a few kilocycles to either side of the center frequency. All transmissions remain within the pass band of the air-borne receivers. Since aircraft at higher altitudes may receive signals from one or more ground transmitters at a time, and frequently do, care is taken to assure audio modulation of the several transmitters in phase, precision control of frequency in the ground transmitters, and the careful design of aircraft receivers to provide sufficient attenuation of objectionable heterodynes falling within the audible range.

50. Selective Signaling. Aeronautical stations are operated on the network principle, using the radio circuit as a party line. Several stations geographically spaced along the flight route share a common frequency. Circuit discipline is accomplished by "listening before talking." Coordination of flight coverage by ground stations is accomplished by intercepting and relaying essential reports. Calls are directed to the desired flight or ground facility by so designating as a prefix to the message. The aircraft crews as well as the ground operator are forced to listen continuously to be assured they respond to all calls directed to them. Considerable valuable information

is intercepted through this monitoring process, but this is becoming less important as the pilot is becoming more self-sufficient through information continuously available to him by electronic aids and the air-traffic-control system. However, the communication function is still essential and in the course of a long flight, particularly when h-f circuits are used, pilot or operator fatigue can cause missed messages, particularly when the signals are weak and nearly lost in noise. If the same principle used in the telephone were applied, so that a particular aircraft need not listen to all communications on a particular channel but could wait until his particular aircraft is discretely contacted, a great deal of the unproductive listening could be avoided. A number of different calling methods have been developed to provide a discrete address or a call-up signal to particular aircraft.

51. SELCAL. One system going into wide implementation among the air carriers is a system known as SELCAL. Primarily used in h-f circuits at this time, although later application to v.h.f. may follow, this system utilizes discrete tones as a calling signal. Two pulses made up of two audio tones each are transmitted sequentially. The air-borne receiving equipment is fitted with reeds which respond only to the specific tone-frequency group to which the aircraft is set. By using the combination of two pulses, each made up of two separate tones, nearly 3,000 high-quality tone groups can be attained from a basic system using 12 tones. By assigning a tone group to a specific aircraft, that aircraft need no longer listen continually on a specific channel but can wait until the SELCAL signal lights a signal lamp or sounds an audible tone in the aircraft cockpit. The SELCAL system has already proved itself in long-haul operation on h-f and promises useful application on v-h-f services as well.

Many different methods of selective signaling are possible and can be designed to provide almost any number of different "addresses."

52. Scatter Propagation. While v-h-f and u-h-f transmission is generally considered as limited to line of sight, a method of over-the-horizon, or *forward-scatter*, transmission has recently been developed to provide reliable communications at distances greater than can be expected of normal v-h-f or u-h-f signals. Inhomogeneities in the troposphere, caused by irregularities in the spatial distribution of electron density, cause the so-called scatter which permits transhorizon communication. In v-h-f scatter systems the signal received over a scatter circuit is of low level although transmission is accomplished at relatively high power. Large directive antennas are used along with very sensitive receivers. Since "scatter" propagation via the ionosphere or troposphere paths is characteristically persistent, improved reliability should be attainable, as opposed to high-frequency communication systems in which frequent and rapid frequency changes are necessary to assure circuit continuity.

Transhorizon circuits are thought to be practical for long-distance transmission of both voice and teletype. They show good promise for voice communication although meteoric interference and flutter caused by signal reflection or aurora may limit the transmission quality. From a propagation standpoint, it is believed feasible that teletype transmission of very high reliability will be obtained from transhorizon systems, although very high power levels are required.

53. High-speed Automatic Communications. As air traffic increases, the need for communication rises as well. In preceding sections, the program for direct center-to-pilot communications has been reviewed and several means for obtaining more radio channels have been described. More and more communication channels and more and more talk over these channels provide for more rapid access from aircraft to the ground controller and from the ground controller to the specific aircraft under his control. However, as more aircraft operate within a terminal area, for example, the need for frequent information on the aircraft position becomes greater. While the situation is helped by providing more communication channels to serve individual control functions, the pilot of a modern aircraft must very frequently change frequencies as he passes through various control areas. It will be seen from later discussion that the growing communication load can introduce delays in the control of air traffic. It is

clear from traffic analyses at particular airports that the message capacity of radio-telephone of somewhere around 200 wpm is too slow to cope adequately with the growing communications load.

The communications load increases at a much higher rate than the increase in air traffic itself, so that limited communications capability will specifically limit the ability of traffic controllers to control flight movements.

In the terminal area the information transmitted during approach and landing involves only a restricted vocabulary. Nearly one-half the talking is concerned merely with the routine mechanics of keeping the process in operation, such things as identifications, acknowledgments, and tests. The remaining time is spent in transmitting information essential to control, such as clearances, position, time, and weather.

One of the major elements of the communications between controller and pilot is the transmission of position reports from air to ground and the resulting control instruction developed by the controller on the basis of these position reports and those from other aircraft in the vicinity. The continuing rapid expansion of air traffic has prompted study and development of automatic high-speed communications, both ground to air and air to ground. While such systems are essential in the military service for the rapid scrambling of military combat aircraft, high-speed communication and rapid access to a large number of aircraft using our domestic airways are becoming increasingly important.

54. Ground-Air Communications. For ground-air communication, the controller must be able rapidly to address a specific aircraft, transmit his message, and receive an answer or an acknowledgment. Much of the ground-air communications is formalized, the variation between messages often being only one number, such as altitude. In the instruction to orbit over a fix, the location designator and altitude over the "fix" may be the only variables. It appears that a high percentage of any ground-air message is routine, although only a few complete messages are routine in their entirety.

A number of different ground-air high-speed communication systems have been considered. A typical one is a digital ground-air link in which advantage is taken of the fact that the total information content of typical ground-air messages is very low. Using a binary system similar to that used in conventional teletype circuits, relatively few bits of information are needed to transmit ground-air control messages. Thus, relatively small band width is needed for this type of system and the message rate can be quite high.

Another of many possible approaches to high-speed communication is a slow-speed television system. In this arrangement, the "camera" is focused on a display of information intended for a particular aircraft and this picture is transmitted over a relatively narrow band width to a receiver perhaps already in the aircraft and is displayed to the pilot. This system has the inherent advantage of the redundancy implicit in the transmission of a picture and permits a visual check by the pilot of the discrete address which can be a part of the transmitted picture.

55. Air-Ground Communications. An air-ground message suitable for high-speed automatic transmission might consist of a position report derived from the aircraft navigation instruments. For example, if both azimuth and distance information is available in the cockpit from ground aids or other instruments, and altitude information is also available to the pilot, a telemetering system might be utilized to send this information automatically air-to-ground upon receipt of an automatic ground-air message. Any of several practical telemetering systems might be utilized for the air-ground portion of an automatic communication system which would, as described here, fall in the category of an air-traffic-control signaling system. Obviously, many items of information might be transmitted air-ground and ground-air to simplify and expedite the air-traffic-control job. The primary requirements for this kind of signaling system are extreme reliability, relatively small band width, and if possible the capability of utilizing transmitting and receiving equipment which already exists in the aircraft.

A high-speed air-ground-air communications system (AGACS) provides three basic elements of information handling:

1. *The Discrete Address.* This is the means whereby the ground operator can, by initiating an action or by a routinely programmed method, contact a specific aircraft by in-flight number or other designator. The discrete address should be automatically checked either by a redundancy check back to the ground or by sending the message address in such a fashion that there is no way to mistake it in the cockpit.
2. *The Ground-Air Message.* The message will consist of the salutation, which is a discrete-address display to the pilot, and the actual message content.
3. *The Air-Ground Reply.* The air-ground reply might consist of an automatic transmission of certain aircraft-position information in response to a request made automatically from the ground, as well as certain yes-no replies, such as "acknowledge," "activate," or "reject," to the message sent on the ground-air path.

It is to be noted that, while an automatic high-speed communication system might render very great assistance to the process of air-traffic control by expediting communications, it is unlikely that all ground-air traffic-control communications can be handled by automatic means. Certain nonroutine messages would probably take longer to prepare for automatic transmission than to transmit on the voice channel. Further, special communications and emergency communications would doubtless be retained on a voice channel. One particularly attractive feature of narrow-bandwidth automatic communication systems is that they present the possibility of multiplexing automatic communications on channels now used for voice communications alone.

IMPLEMENTATION OF THE FEDERAL AIRWAYS WITH COMMON-SYSTEM ELEMENTS

As described in Chap. 26 on navigation systems, many different types of aids have been proposed for use in the common system and many others which, although they have certain specific merits, have not been suggested for incorporation in the system of Federal airways. It is to be recalled that the system of Federal airways and the elements of the common system are intended to take care of the communications, navigation, and air-traffic control for both military and civil users but are not intended to take care of special tactical situations. Thus many highly specialized navigation devices of particular use in combat or other military tactical missions are of little concern to the common system except in so far as care must be taken to prevent interference between different systems which might be used for different purposes. The following is a brief description of some of the navigation aids which make up the major elements used in the system of Federal airways and their state of implementation. Although some specific aids have been described in detail above, they are briefly redefined here. Most of the definitions employed are those developed by the Radio Technical Commission for Aeronautics in its Special Committee 47. They are not stated in order of importance.

56. Low- and Medium-frequency Radio Ranges. Low-frequency–medium-frequency four-course radio ranges define four navigational courses which the pilot of an aircraft can locate and fly using a simple radio receiver operating in the 200- to 415-kc band. The courses are defined aurally by an interlocking series of dot-dash (A) and dash-dot (N) signals of equal amplitude. Transmissions also serve the automatic direction finder (ADF) permitting the pilot to fly toward (*i.e.*, "home-on") the station utilizing the visually indicating cockpit instrument. By combining the bearing information with magnetic bearing information, the pilot can determine his navigational line of position.

These stations also broadcast voice and coded identification signals. The voice function is utilized for weather broadcasts, a service which is primarily for aviation but is being utilized widely by nonaviation groups, particularly by farmers. L-f/m-f four-course radio ranges are used to define the airways system of the United States and are in general use throughout the world. Because of service interruptions and deterioration under various static conditions, lack of flexibility in the number of courses provided, and pilot fatigue in the utilization of the service, the system is being operationally superseded by the

VOR. A limited network, however, will probably be operated as a permanent part of the transition-period common system.

The first l-f/m-f radio range was commissioned in 1928. Their number increased rapidly to a maximum of 378 installed stations in 1949. A simple low-cost air-borne receiver is used to receive the course signals and simultaneous voice broadcasts. As of June 30, 1955, there were 346 such ranges in operation.

As outlined above, the system of v-h-f omniranges (VOR) and the ever-growing number of aircraft capable of receiving VOR signals will permit the Federal airways to plan to discontinue use of all continental l-f/m-f ranges by 1959 with the exception of 88 retained for broadcast purposes and limited navigational use.

57. Low- and Medium-frequency Nondirectional Radio Beacon. "This is a general designation for that class of l-f/m-f radio aids to navigation primarily intended for use with mobile direction finders to provide bearing information.

"The *H Facility* is a nondirectional radio beacon operating within the 200 to 400 kc band installed in en route areas to provide navigational guidance, or in the vicinity of an airport as an aid to approach.

"The *Compass Locator* is a nondirectional radio beacon installed at or in the vicinity of the outer marker or the middle marker of the ILS."

The radio beacon signal is received by an air-borne direction finder which shows the pilot the direction (azimuth) from which the ground signal is coming. By using two or more beacons it is possible to establish a fix.

The nondirectional radio beacon was first established by the Bureau of Lighthouses for marine use in 1921. In 1934, a lightweight manual direction finder was developed for aircraft use and shortly thereafter ADF was developed and the air-borne equipments now number into the thousands. There were, as of June 30, 1955, 173 l-f/m-f nondirectional radio beacons in operation.

58. V-h-f Omnirange (VOR). "VOR is the present standard short distance navigation aid facility. It has received international acceptance; a specification is contained in Chapter 3, Annex 10, Standards and Recommended Practices, ICAO, May, 1950.

"The term 'omni-directional radio range' has been used to designate a variety of developments during the past 20 years. It connotes a single-site transmitting facility which radiates a directional signal rotated through 360 deg in the horizontal plane. The character of the signal varies as the signal rotates. The receiving equipment interprets this variable signal in terms of the bearing of the receiver from the signal source."

The development of 100-Mc VOR was initiated in 1937 by CAA. The installations commenced on the Federal airways in 1946. As of June 30, 1955, there were 410 omniranges in operation with an additional 88 facilities in various stages of planning, construction, and test.

59. VORTAC. Because of a specific military requirement for an azimuth system other than the VOR, the VORTAC system has been agreed to by the Air Coordinating Committee as an element of the common system. The VORTAC system consists of a single station which renders three separate services. The first of these is the standard CAA-VOR. In addition, and co-located, is to be a TACAN (Tactical Air Navigation) facility which renders both u-h-f azimuth information and u-h-f distance-measuring service. The plan is for the military users of the system of Federal airways to utilize the TACAN azimuth along with the TACAN distance-measuring service, and for the civil users to use the VOR azimuth service and the TACAN distance, with a possible later transition away from either one of the two azimuth systems.

In the distance-measuring system an air-borne interrogator equipment transmits a signal which is received on the ground. The received signal automatically initiates a ground-station transmission. Upon receipt of this transmission the air-borne equipment measures the elapsed time for the cycle and converts it into a visual indication showing the number of miles from the aircraft to the specific ground facility.

A pulse-multiplex distance-measuring equipment (DME), different from the TACAN distance system primarily in the manner of channeling, had been proposed by CAA and had received international endorsement by ICAO. Because this system is not compatible with the military clear-channel TACAN azimuth and distance system and because of the stated military requirement for the TACAN system, DME stations which had been commissioned for use with VORs were designated to be replaced by the TACAN distance service. Although the CAA pulse-multiplex DME facilities will remain in operation for several years while the transition takes place, it is expected that the pulse-multiplex DME service will be discontinued as the implementation of TACAN systems goes forward to complete the VORTAC system installations.

60. Instrument-landing System (ILS). "Although the designation 'ILS' is an abbreviation of 'Instrument Landing System' the term ILS had been generally accepted to designate the specific system of electronics aid to approach comprising:

"*a.* A *Localizer*, operating within the 108 to 112 Mc band.
"*b.* A *Glide Slope* facility, operating within 328.6 to 335.4 Mc band.
"*c.* *ILS Markers*, operating at 75 Mc.

"As a general term, 'Instrument Landing Systems' is applicable to any type of radio system which employs ground transmitters and fixed directive antennas to define a beam which laterally localizes the runway extension and defines a *slope* plane (as some angle between 2 to 5 deg) leading to the optimum point of touchdown on the runway.

"The Localizer designates the electronic ground element of ILS which provides lateral guidance to a pilot during his approach. Certain instrument landing systems other than ILS also employ electronic elements which give lateral guidance to the pilots during an approach. While such elements may properly be designated localizers this designation should not be employed without suitable modifiers to distinguish such elements from the Common System localizers.

"The Glide Slope facility designates the electronic ground element of ILS which provides vertical guidance to a pilot during an approach."

ILS markers are 75-Mc radio navigational stations (sometimes referred to as *marker beacons*) the signals of which define specific areas along the localizer course line. ILS markers radiate a fan-shaped field pattern in the vertical plane. The major axis of the pattern is normal to the line of flight. The horizontal cross section of the pattern is approximately elliptical. Markers operate a light in the aircraft cockpit to emphasize specific distances from the approach end of the runway.

ILS enables the pilot to make landings at an airport under adverse weather conditions. The precise guidance of the system not only provides safety but expedites the movement of air traffic in the area by reducing the number of missed approaches. Its use provides lower landing minimums and therefore regularity of operations is increased.

The first commercial models were installed in Indianapolis in 1939. As of June 30, 1955, the total of 170 had been programmed with 157 commissioned. The ILS markers, two of which operate as *outer marker* and *middle marker*, are used to define specific locations on the localizer course.

Other uses of the 75-Mc marker system include the *Z marker*, which is a 75-Mc radionavigation station located at an airway l-f/m-f four-course radio range to indicate position above such a range.

61. Other Major Navigation and Traffic-control Aids. A variety of other aids are in use or are coming into use for long-distance navigation. Among others are Consol, loran, and Decca (see Chap. 26). Self-contained aids, using either velocity information derived from Doppler shift measurements or acceleration measurements by inertial means, are coming into use to provide navigation information. With appropriate computers such devices may permit aircraft to fly long distances without reference to ground-based aids, except for periodic correcting fixes. Several other aids, while they basically render navigation information, provide it to the ground station

rather than to the aircraft and thus may become tools of air-traffic control. These aids, including the airport-surveillance radar, precision-approach radar, ground radio direction-finding equipment, airport surface detection equipment, and secondary radar, may be considered primarily aids to traffic control. These are discussed below.

AIR-TRAFFIC CONTROL

Perhaps the most vexing and perplexing problem in aviation is the problem of air-traffic control. The U.S. Air Traffic Control Service is the responsibility of the CAA. The CAA defines air-traffic control as "a service to promote the safe, orderly and expeditious flow of air traffic. In controlled air space the service is furnished both civil and military aircraft. It consists of Airport Traffic Control (control towers), Air Route Traffic Control (ARTC centers), and Approach Control. Approach control is exercised by centers or delegated to towers. From the use of radar has evolved a facility which is named for its original function, a RAPCON (radar approach control), but which is taking on certain air route or area control. Long range radar however is the primary en-route radar facility."

Control of air traffic is not dissimilar in object from the control of ground vehicular traffic but it is tremendously complicated by several significant differences:

1. Air-traffic control requires the surveillance of three dimensions of space rather than two.
2. Where ground traffic requires as its premise the principle of "see and be seen," air traffic control must be exercised in periods when the visibility is good (and one aircraft is capable of seeing other air traffic) but also in foul weather conditions when the visibility is poor and other traffic as well as landing areas are obscured by clouds, fog, rain, or other impediments to vision.
3. While vehicular traffic is essentially all of the same type, that is, all motor vehicles operate within essentially the same range of speed (say from 5 to 90 mph), in air-traffic control the system accommodates traffic whose speeds encompass several orders of magnitude. For example, an aircraft flying at 70 mph may be in close proximity to an aircraft operating at 1,200 to 1,500 mph. The same control system must accommodate all types of air-borne vehicles.
4. Whereas vehicular traffic is restricted to well-defined traffic lanes—the streets and highways—air traffic is capable of operating anywhere in space (up to certain altitude limits) and the traffic-control system must be flexible enough to accommodate air-borne vehicles traveling on defined routes or airways as well as vehicles, such as military aircraft, which may utilize any track in the sky.

62. United States Policy. After long discussion and much soul-searching, it has been agreed that, to be an effective, economical, and satisfactory way of providing for safe and expeditious flow of air traffic, the air-traffic-control system must be a single service which serves all segments of aviation.

The policy of the United States, as pronounced by the President's Air Coordinating Committee in May of 1954, states that it shall be the continuing policy of the United States to

"1. Provide for a single national common civil-military system of air navigation and air traffic control. The national integrated system shall satisfy the air navigation and air traffic control requirements of all civil and air operations except for those special military requirements peculiar to air warfare.

"2. The common system shall be capable of immediate integration with the air defense system of the United States and will constitute an auxiliary to the air defense network.

"3. To provide for an accelerated joint civil military program of research and development to bring and keep the system abreast of the current and foreseeable future operational requirements.

"4. To accelerate the transition to the most advanced concepts of the common system."

Because of the vastly different needs of the various elements of aviation, control of air traffic is a subject which is in continual controversy. Elements of the air-traffic-control system which may admirably serve one segment of aviation are of little or no

use to another. Equipment which involves vast expenditures may be justifiable for one group and totally unattainable to another. It is thus the function of the common-system air-traffic-control development to render a system usable by all segments of aviation.

63. The Function of Air-traffic Control. An air-traffic-control system has been defined by C. C. E. Bellringer, of the International Air Transport Association, as a "closed circuit servo or nervous system wherein the rate of appreciation of a situation change coupled with the time required to take appropriate action must be greater than the rate of change of the actual situation." Stated more simply, although not nearly so well, the following situation might be visualized:

Assume a situation in the vicinity of an airport where a large number of aircraft are deployed.

Some of these aircraft are holding at specific fix points and different altitudes awaiting instructions to land.

Other aircraft, having just taken off, are in the same general area and are climbing from the airport to their assigned altitudes.

Yet other aircraft are overflying the control zone; that is, they have the intention neither to land nor to take off from the airport in question but are merely traversing the airspace to get from one point to another and their path leads them in the vicinity of the airport-traffic-control zone.

These aircraft may be flying in any one of a number of different directions and may be climbing or descending at any particular moment. Additional aircraft from a nearby military base may be practicing scramble and intercept missions and weave in and out of the traffic to simulate the identification of a potential intruder.

Other aircraft such as helicopters may be flying at low altitudes to deliver mail from the airport to the metropolitan area.

It is the function of the air-traffic-control system to be continuously aware of the position of each of the aircraft in the control zone and to plan and provide instructions to the aircraft under control to assure the safe and expeditious flow of traffic. Furthermore, the element of time is vital. The rate of landing and take-off determines the economic utility of air traffic. Economy of time constitutes a major advantage of air travel. Another complication is that the controller's hand is forced by the different classes of aircraft attempting to take off and land at a particular time. Jet aircraft in particular, consuming tremendous amounts of fuel, probably cannot economically tolerate long delays in obtaining permission to take off or land at a given airport. Delays in air-traffic control are cumulative; that is, in a busy area a delay once incurred is difficult to overcome, since traffic tends to stack up, thus complicating the air-traffic-control problem even further.

Thus the job of the air-traffic control is to provide for:

1. An appreciation of the situation as it appears at a particular moment and an appreciation of the manner of change of the situation which is continually taking place
2. An analysis of the situation to determine the appropriate course of action for each aircraft within the traffic-control zone
3. The transmission of instructions to each aircraft to permit it to carry out the desired maneuvers
4. An appreciation of the fact that aircraft under control have followed the instructions, thus rendering a new situation to the controller for further analysis and action

In the tremendously complex task of satisfying the requirements of air-traffic control, electronic aids play a large part, particularly in the situation of the busy control zone. It is clear that where traffic is very light the problem of traffic control is simple. Only when many different aircraft of many different sizes and speeds going in many different directions with many different tasks all assemble in the same locality does the problem of controlling the traffic become a tremendously difficult task.

The areas in which electronic aids can be of help to the controller are as follows:

Communications. Obviously the most important part of an air-traffic-control system is the link between the controller and the aircraft under his surveillance. This

is the life line through which all intelligence passes from aircraft to controller and back to the aircraft. This need not be voice communications but can be automatic transmission and symbolic display of specific information to the aircraft and the controller. Not only is communications important for coordination between the controller and the aircraft but it is equally important to provide the intelligence link between two controllers responsible for adjacent segments of a traffic-control zone.

Location of the Aircraft. While the communications link can provide information on the position and plan of each individual aircraft in a control zone, this information is usually based on the best estimates available from the pilot, who derives the information from his navigational aids or, under visual-flight rules (VFR), from direct observation of things below him. Unfortunately the information available to the pilot is not always accurate enough or complete, so that a method which can give a controller information concerning aircraft whereabouts independent of communication with the aircraft is highly desirable, not only to obtain accuracy on fix for the aircraft but also to correlate the movements of several aircraft in a particular area.

Identification. Another element is the process of identification, whereby a controller needs to correlate an aircraft at a specific location, which he might see on a radar ground display, to a particular aircraft number. Since radar and some of the other position-fixing devices are essentially anonymous, specific identification of a particular aircraft is necessary. The link between an anonymous aircraft and its specific identification as to air frame, aircraft class, and flight plan must be accomplished either by the aircraft performing the identifying maneuvers in response to request and identifiable by the ground operator or by electronic-identification means.

Planning. After information as to aircraft identity and position has been gathered the controller must correlate and formulate plans for the number of different aircraft within a given area. Again, where only a few aircraft are involved a controller might be able to keep the information on aircraft position and their progress in his head and provide air-control service in accordance with his intuitive reasoning. As the number, type, and missions of the aircraft within a particular control zone increase means must be found to perform some of the computations automatically so as to free the controller to make decisions based on the changing situation and give instructions accordingly, rather than require him to spend a great percentage of his time doing the mechanical and routine job of following aircraft positions and movements.

Automatic Communication. Means must be found whereby the communications air to ground and ground to air can be simplified and automatized where air traffic is so heavy as to overburden available communications channels or communication time in continual position reports and movement instructions.

Coordination. Since more than one controller is involved in the movement of air traffic in a large and heavily traveled area, electronic or mechanical means must be found to meet the problem of coordination between controllers so that passing an aircraft from one controller to another can be simplified and speeded. Not only must an aircraft position be passed from one controller to another but also its identification and intended flight path.

Each of the elements described above is essential if the requirement for expeditious traffic control is to be fulfilled as traffic becomes heavier and heavier.

ELECTRONIC AIDS TO AIR-TRAFFIC CONTROL

64. Communications Devices. The communications system described in a previous section forms the means for the controller to instruct the aircraft and the way in which the pilot of an aircraft can render information as to his position and intent within the air-traffic-control zones, for either approach control or en route control. Because of the time required for a communications sequence to take place, high-speed communications such as described in the previous section, AGACS, or other methods of high-speed communications will become more essential as traffic increases. It is to be remembered that, since the process of air traffic control is essentially a series of actions taken in appreciation of a rate of change of the aircraft situation, the amount of communica-

tions required will increase with the number of aircraft being handled but not in a 1:1 ratio. Obviously, as coordination becomes more complex, with more and more aircraft in a particular approach control zone or in a particular route segment, the amount of communications exchange will increase rapidly and in a relationship much more steep than linear. While the addition of new communications channels helps the situation by providing more communications paths, the amount of work per controller increases with the amount of traffic. His ability to communicate by voice with a number of aircraft is limited not only by his own ability for comprehension of the situation but by the access and time requirements for each communications sequence. Thus, high-speed automatic communications, at least for the exchange of routine information, seems to be an essential as the requirements for expeditious air-traffic control become more and more stringent.

65. Location and Identification Aids. Traditionally, the process of air-traffic control has utilized reports of aircraft over fixes. The "fix" is an aircraft position at a specified point over which an aircraft reports, for air-traffic-control purposes, usually identified by a navigation aid. By reporting aircraft over fixes at various altitudes each aircraft in a control situation is assigned a block of airspace more or less in the form of a cocoon, the cocoon being formed by an altitude layer above and below the aircraft's desired altitude, by a lateral separation between aircraft, usually developed from ground navigation aids, and by longitudinal separation between aircraft in terms of time. In other words, air-traffic control is accomplished by moving the airspace blocks or cocoons in whose center an aircraft travels, by separation of altitude, lateral space, distance, and time. After an aircraft vacates a given fix, another aircraft can be moved into the same block of airspace without danger of conflict. The size of the airspace block depends on the quality of the navigation aids by which an aircraft maintains its position within the block and on the rapidity of the ground controller's appreciation of the rate of change of the momentary situation.

If the controller is provided with positive information as to the position of an aircraft, the separation between aircraft both en route and at an intersection can be substantially reduced. The device which provides the controller information as to the position of aircraft in the vicinity of an airport is the airport surveillance radar.

66. Airport Surveillance Radar. This is an existing electronic equipment designed to detect the presence of aircraft within range of the airport. The position of the aircraft is shown with a high degree of accuracy in range and azimuth. Elevation is not shown.

Airport surveillance radar is probably the most important contribution to the expeditious movement of air traffic since the Second World War. By permitting accurate position information to be displayed to the controller, the utilization of the airspace in the vicinity of an airport is materially increased. For example, at an airway intersection where, without radar, a 10-min period is required before an aircraft is permitted to cross the path of another aircraft, with radar, separations of 1 min are permissible. In the case of an aircraft traveling at 300 mph, this amounts to a difference in the required distance separation of 50 miles when no radar is used and 5 miles when radar is used. Comparable reductions in time and distance separation are attained in the en route arrival and departure situations. The substitution of positive 3-mile separation for the time or altitude separation required in a nonradar operation has permitted a considerable reduction in the interval between successive landings of aircraft on the runway and successive departures of aircraft from the runway. This results in a corresponding reduction of delays of aircraft under high-density traffic conditions.

The airport surveillance radar is an outgrowth of the radar approach and landing system known as GCA developed during the Second World War. The CAA has installed a total of 31 airport surveillance radars in major traffic hubs with plans for additional equipments under way.

67. Long-range Radar. While the airport surveillance radar is an excellent air-traffic-control aid for the approach control zone, it does not provide sufficient range to

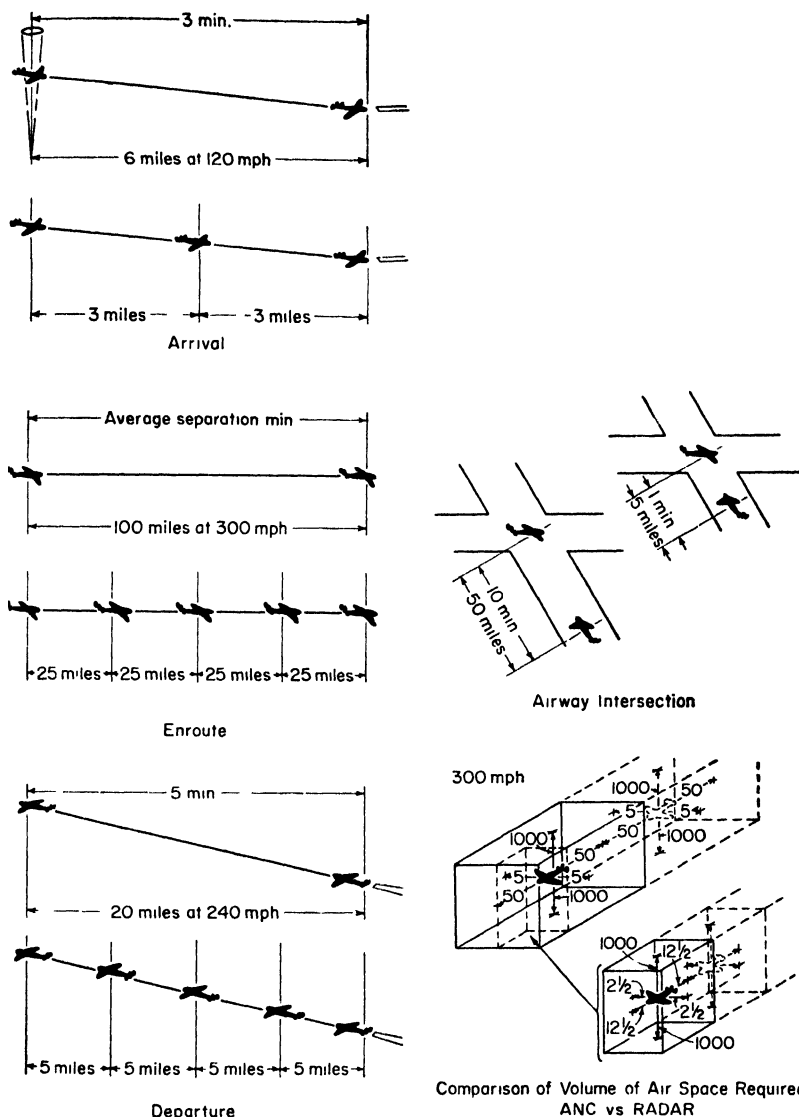



FIG. 7. Manner in which radar increases airspace utilization.

permit its use in the en route control of air traffic under instrument flight rules (IFR). To avoid stacking up of aircraft in the vicinity of airports, flow control of air traffic reaching an airport must be instituted while the aircraft are still far from their destination. As an aid toward the control of traffic, the long-range radar promises to provide excellent service as an en route aid as well as approach control aid. Long-range radar is not only useful in the control of air traffic but is also a vital part of our air-defense system. As a result, it is planned that integrated use will be made of long-range radars

for both air-traffic control and air defense, even though not all defense radars are technically and operationally suitable for joint use.

68. Secondary Radar. While the use of radar as an aid to air-traffic control has influenced aviation tremendously the basic radar has one major drawback. Each aircraft looks like every other aircraft, and thus there is almost total anonymity about the blips to the air-traffic controller. Since it is necessary for the controller not only to be aware of an aircraft in a particular location, but also to know the identity, type, and flight plan for that aircraft, identification becomes almost an essential. Secondary radar, or the ATC radar-beacon system, solves the problem of identification of the individual aircraft in air-traffic control.

The secondary radar is an electronic means of detecting and identifying aircraft. Whereas primary radar relies on the reflection or echo of ground-transmitted radar signals from the surface of an aircraft structure, secondary radar utilizes the transmission of radar replies from the aircraft. The ground station sends up a coded interrogation at one frequency (1,030 Mc) to the aircraft. Aircraft equipped with an airborne ATC transponder reply on a different frequency (1,090 Mc) with pulse-coded reply signals. By utilizing between two and eight reply pulses in the reply group, specific binary code identification is attained. By decoding the replies on the ground and displaying symbolic information synchronously with radar echo, specific identification codes, assigned to aircraft by the air-traffic controller, can be displayed in any number of ways on the ground controller's display, thus lending specific identity to the previously anonymous blips. By assigning identification codes either by aircraft type, by route, by holding fix, by altitude, or by any one of many different methods, a considerable amount of information about an aircraft can be transmitted via this identification link, without requiring continuous communications between the aircraft and the traffic-control center. The ATC radar-beacon system has obvious uses also in identification of aircraft for purposes of air defense, and promises to eliminate the need for some of the identification maneuvers now required of aircraft entering air-defense zones within the United States.

The ATC radar-beacon system, in addition to providing the basic identification feature, will reinforce weak primary-radar signal returns from jet fighter and other types of aircraft having small reflection areas. It will serve the further function of providing strong returns under conditions of heavy precipitation which render the present-day S-band surveillance radar system virtually useless. It is the CAA plan to install ground ATC interrogators as an adjunct to all airport surveillance and long-range radars, including a portion of the military radars planned for joint air-defense–air-traffic-control use.

69. Precision Approach Radar (PAR). As an aid in the final approach of aircraft to land, the precision approach radar is a system designed to monitor the progress of an aircraft to the instrument runway and to present the information to the controller in the tower with a high degree of accuracy. Distance, bearing, and elevation above ground are presented continuously on the face of the radar scope. PAR in conjunction with airport surveillance radar constitutes the military ground-controlled approach (GCA) whereby a controller can "talk" the pilot down to a successful landing. In civil use where ILS is widely used and where the pilot himself makes his landing on the basis of information from the electronic glide-slope and localizer beams, the PAR facility is used primarily by the air-traffic-control agency to monitor approaches and to warn the pilot if he is too low, too high, or off the center line of the instrument runway.

PAR equipment is an outgrowth of the military radar developments during the Second World War. A total of 10 precision approach radars have been installed at major traffic hubs as of June 30, 1955.

Since PAR is used only in the approach area leading to the instrument runway, it scans only a small portion of the total airspace around an airport. Present-day precision approach radars look only 20 deg in azimuth and 6 deg in elevation. These triangular segments are expanded in presentation and occupy almost the full area of

the radar controllers' display. With PAR operated in the 3-cm band, the position of an approaching aircraft may be determined within about 20 ft of elevation, 40 ft in azimuth, and 300 ft in range when the aircraft is a mile from the end of the runway.

70. Airport Surface-detection Equipment (ASDE). As our airports grow larger and the runways longer, air-traffic control after touchdown becomes an important part of the air-traffic-control job. Accurate information to the tower and operator as to the progress of an aircraft along the runway during bad weather can permit more frequent landings and take-offs than if no visual surveillance were possible. It is an acute problem at high-density airports because runways and taxi strips which may be nearly out of vision of the controller even on clear days must be immediately known clear if traffic is to be moved without serious delays.

ASDE is a radar usually operated in *K* band which looks directly at the airport and has very short range. Its only purpose is to display traffic on the ground on runways or taxi strips. The provision of ASDE is of primary use in large heavy-traffic metropolitan airports, and it is planned that approximately 20 locations will be equipped within the next several years.

ANALYSIS OF POSITION AND PROGRESS OF AIRCRAFT

By the use of navigation aids for the aircraft and information on position and progress of the aircraft made available to the ground controller through radar, secondary radar, and other devices, we can provide nearly all the tools for the control of air traffic. The only problem is that in a high-traffic situation information comes to the controller so fast and from so many different sources that rapid analysis of the constantly changing complex situation is a formidable task. Therefore, it will become necessary not only to automatize the development of information on aircraft position and progress, but also to automatize the computations which must be accomplished before a controller can make decisions in the traffic-control situation.

71. Computer Applications. A large number of developments, both military and civil, are working toward the goal of simplifying the controller's task. The work is predominantly in the field of air-traffic-control computers which store and compute the information about specific aircraft and their progress. By feeding into the computer the present position of all aircraft under control, their speed, and their course, a computer can store this information and compute the situation that will exist a few seconds or a few minutes later. Depending on the sophistication of the computer and on the method of feeding data to it, the computer either can store the information and give it back to the controller on demand or can continually compute the impending situation by means of flight plans fed to it or by continuous information fed to it from radar sources or by the controller himself, and thus warn of impending conflicts as well as provide the controller with instructions for various aircraft to provide satisfactory control. It is apparent that, while the individual acts of a computer for traffic control, for example, are relatively simple arithmetic computations, the traffic-handling capacity required of such a computer and the absolute reliability required make automation of the traffic-control process a complex procedure.

The possibilities of computers improving our air-traffic-control systems are virtually limitless, and discussion of the many detailed systems is beyond the scope of this book. It may be useful, however, to suggest briefly two areas in which considerable work is going forward.

One device is a storage system for flight-plan data, basically a storage drum, which is fed information from teletype-machine sources, and can serve to store flight plans which might have been filed by the aircraft which are to come under control in a particular control area. Fed flight-plan information and progress reports periodically, the equipment can advise the controller, on the basis of time, of traffic which will come into his control at a particular time and in a particular manner. An extension of this arrangement would feed flight-plan corrections into the drum as they are received, so that only up-to-date information is displayed, and add computers which would follow flight progress automatically and warn of impending conflicts.

This type of traffic-control-coordinating equipment is only one of a number of types intended to simplify the controller's task by automatically providing him with information which he would normally compute laboriously or sort out of a maze of individual communications.

On the more complex side of the picture is a system developed for military use called the semiautomatic ground environment (SAGE). This system, while not specifically intended for civil air-traffic control, illustrates a class of system where such control might be automatized almost completely. In this system, the basic source of information is a series of ground radars. The ground equipment automatically performs various computations for the purpose of guiding interceptor aircraft and in aiming other weapons. It is capable of combining the coordinates of several radar sets one with another into a common display system and it acts as a communication network between the operators.

This system is intended to provide a central machine with a large and rapidly functioning memory whose job would be to keep track of all aircraft. All data are sent to this machine automatically, and the function of men in the data-transmission process would not be to transmit data but to inhibit the data when they obviously become false. This system requires not only a number of radars to cover a particular area but a large network of communications to feed the automatic machine.

In the study of air-traffic control and the electronic devices, from radars to computers, which might be used in effecting air-traffic control, it has become apparent that we have at our disposal the scientific tools to do almost any computation and the storage of any information we might see fit. One important consideration, however, is that of economics, for while nearly any job can be done, the complexity and expense of providing such service may readily become prohibitive and thus unattainable economically.

A further consideration in the automation of an air-traffic-control system which has troubled system designers for many years is the basic question of reliability of automatic systems. While some argue that an automatic system is not subject to human error, and should be far more reliable than the equivalent human controllers, others point out that no automatic system, no matter how sophisticated, can exercise judgment in emergency situations. The latter group feels, therefore, that while we should make every possible use of automatic devices to simplify the job of the air-traffic controller, we should always let the decision-making authority remain with the human controller who has the judgment to take over the situation where the computing machinery has failed.

AIR-TRAFFIC CONTROL AND THE VARIOUS CLASSES OF AIRCRAFT OPERATORS

In the previous sections of this discussion, a number of different air-borne equipments have been discussed. It is important to note that the Federal airways air-traffic-control system must accommodate not only the aircraft equipped with the latest and best of air-borne aids but must also take care of the aircraft which cannot, for any of many reasons, carry sophisticated equipment. Whereas modern airliners might carry all possible equipment which could be useful in air-traffic control, many private aircraft will have to make do with a minimum of air-borne equipment. Thus, any modern air-traffic-control system must be capable of accommodating all users of the airways, those with a minimum amount of equipment and those with complex and sophisticated electronic installations.

THE PECULIARITIES OF AIR-BORNE ELECTRONICS EQUIPMENT

72. Reliability. Electronic equipment in aircraft must be reliable. Frequent failures may abort an expensive and important mission. Yet while reliability and accuracy in air-borne electronic equipment are essential, the environment aboard an aircraft places nearly every handicap on the attainment of this goal. We require lightweight and small-sized electronic equipment to make it economical to carry, yet we

expect extreme ruggedness from the resulting equipment. The equipment must withstand the shock and vibration encountered during flight and during ground transportation over rough terrain. We subject the equipment to considerable variations in voltage and frequency. We expect the equipment to operate under a very wide range of temperatures and a wide range of altitudes. With all this we expect long-term stability and need of only infrequent checking or calibration of the equipment.

73. Weight, Size, and Cost. There are generally three categories of air-borne electronic equipment. The first of these is the equipment used aboard itinerant aircraft, generally in the smaller single-engine or light twin-engine types. In this type of equipment, size and weight are at a premium, but perhaps cost is even more important. In addition to these restrictions, the small itinerant aircraft usually has only a limited amount of electrical power available, so that low power drain is an additional requirement. Further, the economic situation of many itinerant aircraft operators is such that maintenance of equipment will be often done primarily when something has gone wrong, so that emphasis must be placed on extreme reliability and deteriorative failures or time-dependent loss of calibration must be avoided. Aboard the modern itinerant aircraft, the first requirement is for communications equipment, with simple navigation equipment closely following in importance. Low-cost versions of communications equipment and of some navigation devices have been developed for light-plane users.

At the other end of the scale in complexity and cost is the air-borne equipment installed on military aircraft. Subject to much greater rigors in terms of altitude, shock, and vibration, etc., than light-plane equipment, tremendously greater demands are also made upon it. Because the weight and amount of electronic equipment directly affect the performance and combat capability of military aircraft, extreme emphasis must be placed on miniaturization. It is in the military fields that the new frontiers must be opened not only in the miniaturization of the equipment but also in providing reliable performance. Military aviation electronics equipment must assure long-term reliability and stability for big, long-range aircraft, and must be equally reliable in tiny fighter aircraft where there is little space for the pilot, much less for bulky electronic equipment.

Military electronic equipment is often located in nonpressurized areas and much of the gear must be built in pressurized containers to withstand altitudes up to 75,000 ft, as well as extreme variations in temperature, and not only the shock and vibration of normal aircraft maneuvers but also that resulting from the firing of weapons.

The third category of air-borne electronic equipment is that used by air-carrier aircraft. In this mode of operation size and weight are obviously also important but to a somewhat lesser degree than reliability. To assure reliability of air-borne electronic equipment used in air-carrier service, most devices are carried in duplicate and independent systems are provided to permit easy switching from one to another in the event of failure. Electronic equipment aboard air-carrier aircraft is usually located in a pressurized area near the crew compartment so that it must withstand neither the rigors of extreme altitude variation nor extreme temperature variation often required of military equipment.

On the other hand, whereas military electronic equipment is usually intended to have a useful operating life in the vicinity of 1,000 hr, air-line equipment is required to operate for much longer periods. This disparity arises from the fact that military aircraft, intended for performance of a particular mission in wartime or the training for that particular mission in peacetime, are of necessity in a waiting status a high percentage of time. The air-carrier aircraft, and the military transport service as well, are required to operate 9 to 12 hr per day every day so that 1,000 hr of operating life is usually accumulated in a period of 3 to 4 months. Equipment carried aboard air carriers is normally expected to operate satisfactorily for a minimum of 1,000 hr and usually 1,600 to 2,000 hr between scheduled removals for maintenance and service.

The weight of electronic equipment is a vital concern to the designer and user of

modern aircraft and thus great emphasis is placed on lightweight reliable air-borne electronic equipment. In a modern air-carrier aircraft, there are approximately 26 radio or electronic systems with hundreds of switches, control knobs, push buttons, indicating instruments, etc. This requires approximately 55 cu ft of space and approximately 2,400 lb for the total installation. Each pound of electronic equipment not only increases the basic cost of the aircraft but reduces its payload capacity and increases its cost of operation. Therefore, great attention must be paid to the reduction of weight wherever possible, but without a reduction in over-all reliability. Unreliability can cost many times increased operating cost due to the weight of the electronic equipment.

74. Aircraft Antennas. Whereas the designer of radio equipment or electronic equipment for ground or ship-borne use is usually little concerned with the aerodynamic drag introduced by a fixed antenna, this problem is of paramount importance in design of antennas for aircraft. While the aerodynamic problem of where to locate antennas has been present since the first radio went aboard an aircraft, the problem is severely aggravated on the new high-performance aircraft types. As an example, the air-borne ADF requires two antennas, a loop and a sense antenna. The loop antenna was, and on many aircraft still is, housed in a football-shaped housing which protrudes from the aircraft belly. The sense antenna is a wire antenna suspended away from the aircraft skin on two masts. While on present-day aircraft these antennas cause aerodynamic drag, this drag can be justified on the basis of the performance attained from the ADF system. On high-performance aircraft the costs in terms of aerodynamic drag and loss of aircraft performance make the old type of antenna arrangement totally unusable. Instead, a loop is used which is essentially submerged in the belly of the aircraft, and a sense antenna which is also nearly submerged and is flush with the skin of the aircraft. Quite naturally, as the aerodynamic drag imposed by the ADF antenna system is thus so reduced, the performance may also be reduced, since submerged radiators are often not so good as antennas with more effective electrical height. It has become necessary to improve the electronic equipment itself to the point where the equipment can compensate for the reduced performance received from the antennas.

One other case of interest is the h-f communications-system antenna. For the h-f antenna on conventional present-day aircraft, a long wire is normally used, suspended well away from the aircraft and usually extending from the vertical stabilizer well forward to the nose of the aircraft. On high-performance aircraft, this kind of an antenna may be unacceptable. In its place designers are using an insulated section of either the wing or the vertical stabilizer to serve as an antenna. This is accomplished by electrically isolating the tip of the vertical stabilizer, for example, from the rest of the aircraft and exciting this isolated section.

The operational characteristics of an aircraft at a given speed affect the determination of the tolerable amount of drag due to an antenna. On new jet air-transport aircraft, 1 sq ft of parasitic drag area (roughly equivalent to the area of a flat plate producing an equivalent drag) is equivalent to 1,200 lb of additional weight. The submerged type of antenna, even if it provides good electrical performance, cannot be judged on this criterion alone. In pressurized aircraft intended to travel at high speeds, the submersion of an antenna generally requires considerable structural beefing in the vicinity of the antenna to preserve the integrity of the air frame; thus even the submerged antennas cost a great deal in actual weight. For each instance an analysis must be made to determine the relationship of the structural weight of a submerged antenna against the aerodynamic drag (and thus effective weight) of an antenna which protrudes from the aircraft.

75. Electrical Power. All aircraft are limited in the amount of electrical power available for electronic equipment. Economy in the utilization of electrical power is a requirement for all types of aircraft.

The power sources aboard modern aircraft can be roughly divided into three general types:

A d-c supply is usually 6 or 12 volts for small itinerant aircraft, 27.5 volts nominal for air-carrier and military aircraft. The d-c power supply in most older transport aircraft is the basic source of electrical power and it is generated from an engine-driven generator. Voltage regulation of the 27.5-volt supply of +10 and −20 per cent is normal. Where d.c. is the only source of basic power aboard the aircraft, inverters are widely used to provide the necessary a.c., although many air-borne radio-equipment types, used in both air-carrier and itinerant service, utilize dynamotors and (particularly in the itinerant service) vibrator power supplies to supply high voltage.

The a-c supply in a basically d-c airplane is usually obtained from a 400-cps, 115-volt, three-phase, one-side-grounded inverter. Power from this type of machine may vary in voltage as much as ±10 per cent from the nominal 115 volts and the frequency may vary 20 cps. On some present-day military aircraft, a-c power is available also from inverters which are not regulated to a specific frequency and which may vary from 300 to 1,000 cps, with voltage variations of ±10 per cent.

On new transports and many new military aircraft the power sources are predominantly 400-cps a-c supplies from engine-driven alternators. Considerable weight savings are attained in this type of basic electrical-power supply and close control of frequency and voltage will be available to the air-borne electronic equipment. In the a-c aircraft, the d-c supply, which will continue to be used widely for operation of control circuitry and for some transistor equipment, will be of somewhat poorer quality. This power will be obtained from a transformer-rectifier system with only limited voltage regulation.

With this change in the power-supply concept aboard some military and new air-carrier aircraft, it becomes a requirement for air-borne electronic equipment to provide considerable flexibility in power-supply arrangement. Electronic equipment designed for a particular purpose must usually be adaptable to either a basically a-c or a basically d-c airplane. Both varieties of aircraft will be in operation for many years.

76. Interference and Interference Reduction. The designer of air-borne electronic equipment must accept as one of his design problems the fact that air-borne electronic equipment must live in an electrical environment which often produces interference. He must consider these conditions and design the equipment to perform its normal functions in the presence of interference.

Interference may be incurred from atmospheric disturbances, man-made noise (such as from aircraft electrical equipment or other radio equipment in the aircraft or received at the antenna), or inherent tube noise. An effort to eliminate the effects of interference must be based on where the interference originates. Some forms of interference will originate within the equipment itself or will result directly from other equipment or accessories in or around the aircraft. Other types of interference will be encountered which are due exclusively to effects occurring outside of and at a distance from the aircraft.

The designer of an aircraft electrical system has the responsibility for assuring that an aircraft is essentially noiseless electrically. This means in theory that the aircraft is designed to suppress ignition noise and to assure that power lines circulating within the aircraft do not carry radio and electrical noise. It is difficult for the aircraft designer to provide an aircraft which is completely noise-free, and it is therefore necessary that the electronic equipment be designed to be tolerant of considerable noise and ripple on power control entering the equipment.

One major cause of interference, particularly in the low- and medium-frequency regions, is atmospheric interference due to *precipitation static*. Precipitation static is radio interference experienced when the flight path of the aircraft is such that disruptive static discharges are caused from the aircraft. Precipitation static is developed by flying through snow or dust storms, raindrops, thunderstorms, ice crystals, etc. The elimination of sharp edges or pointed protuberances and the bonding of separated parts of the aircraft structure help to reduce the precipitation-static problem. The use of discharge wicks and improved antenna wire and insulators all appear to help the problem somewhat. Although the smooth external characteristics of high-speed

modern aircraft with essentially dragless antennas should make them more immune to interference from precipitation static, the use of dielectric materials such as Plexiglas and Fiberglas appears to cause serious new precipitation-static effects. Plastic bubble canopies on fighter aircraft appear to be serious offenders in causing precipitation static. While developments go forward toward cleaning up the aircraft to prevent fast static discharges, other developments are working toward the elimination of the effects of the static by circuitry methods, for example, by disabling the air-borne receivers during periods of heavy discharges.

References

"Air Traffic Control and the National Security," Air Traffic Control and Navigation Panel, Air Coordinating Committee, December, 1950.

"Air Traffic Control, Report of RTCA Special Committee 31," RTCA Paper 27-48/DO-12.

"The Air Traffic Story," RTCA SC 47, RTCA Paper 194-52/DO-47, Dec. 1, 1952.

"ARINC Synchro System Manual," Aeronautical Radio, Inc., January, 1954.

"Automatic Flight Control," RTCA Special Committee 18.

"Aviation Services, Rules and Regulations," Title 47, Telecommunications, Chap. I, Part 9, FCC.

Bailey, D. K., *et al.*: A New Kind of Radio Propagation at Very High Frequencies Observable over Long Distances, *Phys. Rev.*, **86** (2), Apr. 15, 1952.

Carnes, W. T., Jr.: ARINC Electronic Equipment and Planning for the Airline Industry, *Trans. IRE*, **PGA-10**, December, 1953.

Civil Air Policy, The President's Air Coordinating Committee, May, 1954.

"Design Techniques for Interference-free Operation of Airborne Electronic Equipment," USAF Wright Air Development Center.

"Electronic Navigation Systems," RTCA Paper 87-51/DO-42, Special Committee 55, Definitions, Air Navigation Terms, Nomenclature.

"Federal Airway Plan," Fiscal Years 1957 to 1961, CAA, U.S. Department of Commerce, December, 1955.

"Federal Airways Air Traffic Activity, Fiscal Year 1955," Civil Aeronautics Administration.

"Guidance for Designers of Airborne Electronic Equipment," Aeronautical Radio, Inc., ARINC Report 403, Sept. 1, 1955.

Hilton, John H.: "Extended Area of Communications," ATC Centers and INSACS, Paper, RTCA Assembly, fall, 1955.

"Instructions for the Operation of Communications Facilities," CAA, Federal Airways Manual of Operation, IIB, Jan. 19, 1950.

Norton, K. A.: Point-to-point Radio Relaying via the Scatter Mode of Tropospheric Propagation, *Nat. Bur. Standards Rept.* 3562, Central Radio Propagation Laboratory.

Page, L.: "Introduction to Theoretical Physics," Van Nostrand, 1928.

Petry, C. A.: "Airlines Communications in the Jet Age," Aeronautical Radio, Inc.

Ringeon, R. M.: VHF Trans-horizon Communications System Design, *IRE Convention Record*, Part VIII, 1956.

Valley, George E., Jr.: Operational Applications of SAGE, spring assembly, RTCA.

CHAPTER 26

AVIGATION ELECTRONICS

By Anthony Casabona*

1. Introduction. The success of any transportation system depends to a large measure upon the rigorous maintenance of safe scheduled operation. During the last decade the airplane has become a most important factor in transportation and it is recognized that radio is an indispensable tool in maintaining and furthering the aviation transport industry. Today the scheduled air lines of the United States average hundreds of millions of miles annually, with air mail and air express being carried in large volumes. In addition, military, business, and private flying has increased at a rapid rate. These factors, coupled with the use of high-speed modern aircraft, have created many problems in the safe conduct of air operations. Such operations are conducted daily around the world and would be virtually impossible without modern radio aids.

Radio provides aviation with communication, navigation during all phases of flight, traffic control, collision-warning services, and airport-surface control. From the time that the first piece of radio equipment was installed on a transport airplane the ultimate goal has been the unification of these basic functions into an integrated system directed toward the continuous and safe movement of air traffic. As the operational problems of air transport have become progressively more complex, there has been a gradual shift of responsibility for aircraft control from the individual pilot to a ground controller. This transition can be expected to continue as air traffic and speeds increase.

In the division of responsibility between ground controllers and crews of aircraft, the ground controllers must be aware of the flight plan to be followed and must make frequent checks of the degree of adherence to the prescribed plan. In the present state of the art, the air crew must be given sufficient information to carry out the prescribed plan with minimum reference to the ground controllers. Ultimately, it appears necessary to by-pass the human navigator and to apply the navigational intelligence directly to the controls of the aircraft. In such a system, the air crew should have means for correcting the flight of the craft if it does not adhere to the control plan, or for deviating from the plan in case of emergency. The ground controller must have a continuous display of position of each craft in flight, with means for issuing control signals to conform to existing traffic requirements and flight conditions. Automatic ground computers, and associated automatic pictorial display of the monitored data, must form an important part of any efficient traffic-control system. Regardless of the computation equipment required to process the data, the input data themselves must be collected by one radio means or another. Accordingly, this has become an important additional requirement on the organization of the basic radio services.

2. Organization of Civil-aviation Radio Facilities of the United States. The Civil Aeronautics Administration (CAA) of the U.S. Department of Commerce is the

* International Telephone and Telegraph Laboratories, formerly Federal Telecommunications Laboratories. The author wishes to express his gratitude to many individuals, and to the management, of the ITT Laboratories for their assistance and cooperation during the writing of this material; to Richard P. Battle, Deputy Chief, Facilities Division of the Civil Aeronautic Administration, New York International Airport, for his review and valuable suggestions concerning the information on CAA facilities; and particularly to Brigadier General Peter C. Sandretto, Vice President and Technical Director of the ITT Laboratories, for his advice and the use of material from his book "Electronic Avigation Engineering."

government organization responsible for the installation, maintenance, and operation of civil-aviation radio facilities, for research and development leading to their improvement, and for the dissemination of information. Installation and maintenance functions, which include the planning, design, construction, installation, and improvement of air-navigational aids, are carried on by the Air Navigation Facilities Service; the actual operation of facilities is carried out by the Airways Operations Service; and research and development activities are conducted by the Technical Development Service.

Airway facilities, exclusive of communications, directly operated by the CAA as of Jan. 1, 1956, are listed in Table 1. A small portion of the VOR, the Federal airways system, is shown in Fig. 1.

Table 1. Extent of CAA Facilities for Air Navigation and Control

Item	Number
1. Number of facilities, as of Jan. 1, 1956	
a. Mileage of civil airways	
(1) Total:	
V-h-f	79,500
L-f/m-f	69,770
(2) Under traffic control	44,057
(3) Oceanic routes	14,400
b. Air-navigation radio aids	
(1) Adcock ranges (RA and MRA)	344
(2) Loop ranges (RL, MRL, and ML)	
(3) V-h-f ranges (VOR)	419
(4) VAR	6
(5) H-type markers (radio beacons on airways)	175
(6) K-type markers (compass locators)	42
(7) Fan markers (full-powered)	236
(8) Fan markers (low-powered)	28
(9) Instrument-landing system (ILS)	157
(10) Approach-search radar (ASR)	31
(11) Precision-approach radar (PAR)	10
(12) Distance-measuring equipment (DME)	234*
c. Traffic control	
(1) Combined station tower	75
(2) Airport traffic-control towers	100
(3) Airway traffic-control centers	31
2. Airway light beacons	842

* DME of this type has been decommissioned in accordance with the Air Coordinating Committee's decision of 1956 and will be replaced by a Tacan-type DME, often referred to as DMET. (See Sec. 7.)

Other agencies of the government have duties requiring the operation of radio facilities which can be utilized by civil aviation. Among these are the Army, Navy, Coast Guard, and the Federal Communications Commission. As a service to pilots, CAA maintains close liaison with these facilities, disseminates up-to-date information on their status or on sources of information regarding their status, and, where needed, establishes services to facilitate their use.

3. Information Sources. Current information on air-navigation radio aids is given in the following publications:

Flight Information Manual. U.S. Department of Commerce, CAA, Office of Aviation Information. Issued semiannually. Includes air-traffic control and communications procedures, instrument-approach procedures, direction-finding station data, airport and radio-facility indexes, foreign-entry and flight requirements, and radio call letters and frequencies.

Airman's Guide. U.S. Department of Commerce, CAA, Office of Civil Aviation. Issued biweekly. Includes information on new facilities, facility changes, and notice to airmen on temporary conditions, such as airway construction or inoperability of facilities.

Radio-facility Charts. U.S. Department of Commerce Coast and Geodetic Survey. Revised weekly. Contains 42 separate charts providing complete radio information to facilitate the planning and execution of cross-country flights in the United States. Shows all radio-facility data with detailed tabulations of all radio stations in operation, classified as to aids to navigation and control towers.

Catalogue of Loran Charts and Service Areas. U.S. Navy Department Hydrographic Office. H.O. Pub. 1-L. Includes description of loran system, discussion of factors affecting

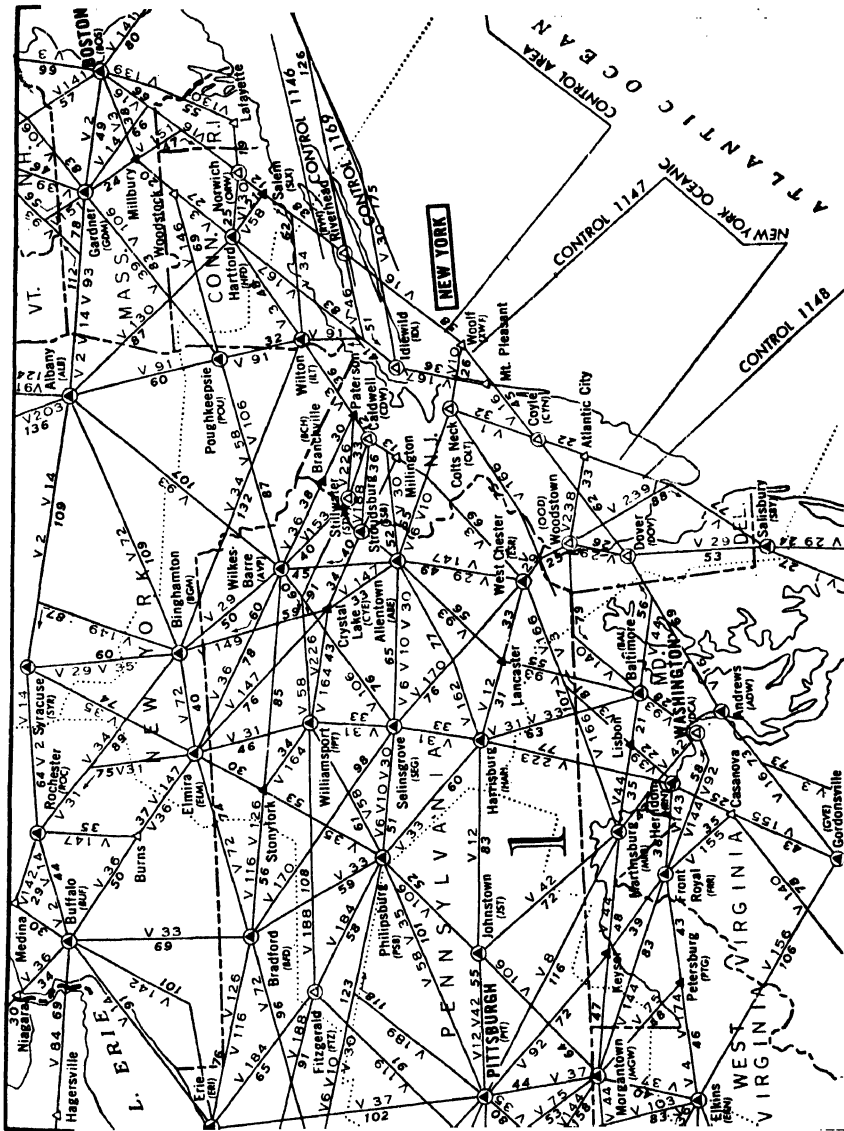



FIG. 1. A portion of the VOR airway system. (*Courtesy of CAA.*)

accuracy, shows ground-wave and sky-wave coverage, and presents a chart index of loran service in the Atlantic and Pacific Oceans.

Army Air Forces Radio-facility Charts. Headquarters, Air Materiel Command, Wright Field, Dayton, Ohio. Issued monthly. Furnishes data and charts for use by Air Force pilots; shows how military aircraft utilize virtually the same radio-navigational facilities as do the civil aircraft.

United States government facilities comprise many diverse types of equipment. The following brief descriptions of these facilities are intended to explain their functions; technical details will be presented in subsequent sections.

4. Radio Ranges. Radio-range stations provide radio-defined courses for delineation of lines of flight during restricted-visibility conditions and over-the-top operations when visual contact with the ground is impossible.

The first type applied to the airways was the l-f directional, four-course, aural radio range. Some 344 are still in use. These operate within the band 200 to 400 kc, with a modulation frequency of 1,020 cps. Each range has four courses, produced by the alternate interlocked Morse-code keying of two figure-of-eight patterns, one N and the other A, directed at approximately right angles to each other, the exact direction of each course being that which will best serve the airway and/or airport. Ranges are normally located within 2 to 4 miles of an airport, at least one course being aligned in such a manner as to facilitate instrument approach to same.

The position of a course is determined by comparing the relative strength of the signals N and A, the proximity of the course being indicated by approaching equality of these two signals. Exactly on course, the interlocking of the two signals produces a steady monotone (1,020 cps), interrupted only for station identification. The latter, consisting of a two- or three-letter combination, is transmitted twice every 35 sec, first in the N quadrant and then in the A quadrant.

Voice transmissions are made on the range frequency unless otherwise indicated by the inclusion of the letter W in the station designation.

These l-f directional ranges can be further classified by type of radiator utilized and by power output as follows:

Tower Radiators. Five steel towers approximately 130 ft in height, four forming a square and one in the center, are utilized for the following designations:

RA, power output between 150 and 400 watts, usable distance 65 to 100 miles
MRA, power output between 50 and 150 watts, usable distance 40 to 75 miles

Loop Radiators. Two crossed, vertical loops mounted on five poles approximately 50 ft in height, four poles forming a square and one in the center, are utilized for the following designations:

RL, power output between 150 and 400 watts, usable distance 50 to 75 miles
MRL, power output between 50 and 150 watts, usable distance 30 to 55 miles
ML, power output less than 50 watts, usable distance 20 to 40 miles

5. V-h-f Directional Range (VAR). This is a two-course visual, two-course aural range with simultaneous voice feature operating within the band of 112 to 118 Mc. The power output is approximately 200 watts, and the usable distance is limited by line of sight, depending upon the altitude of the aircraft and the terrain, with an upper limit of about 200 miles at very high altitudes. The radiators employed are small and closely spaced, being located in a small building mounted on a tower 30 to 60 ft in height; a counterpoise approximately 35 ft in diameter is also mounted on the tower.

The visual courses of VAR are 180 deg apart and are created by the overlapping of two field patterns essentially cardioid in shape, one modulated at 90 cps and the other at 150 cps. An electrical indicating zero-center instrument, actuated by a receiver designed for the purpose, is utilized in the aircraft to locate and fly the courses. This is accomplished by comparing the relative strengths of the two opposite sectors. The indicating pointer, or needle, is centered (vertically) on the instrument scale when the aircraft is exactly on course. Full-scale deflection of the needle will occur at approximately 10 deg off course.

The aural courses of VAR are also 180 deg apart and are located at right angles to the visual courses. The aural courses are produced by the alternate interlocked Morse-code keying of the letters N and A in a manner similar to that of the l-f directional range, except that the field patterns are like those utilized for the visual courses of VAR. The modulation frequency of the aural signal is 1,020 cps, and like the l-f directional ranges, the aural signal is interrupted approximately twice a minute for transmission of the station identification. The orientation of the N and A sectors with respect to the visual courses is always the same; *i.e.*, N will be on the east leg of a visual east-west course or on the north leg of a visual north-south course.

6. Omnidirectional Radio Range (VOR). The omnidirectional radio range radiates a signal which, through the use of special receiving equipment, may be utilized to establish a visual course in any radial direction with respect to the transmitting station. Such a station may be considered to have an indefinite number of flyable courses. It was developed to provide the type of service required by the concept of the universal airway. An "azimuth-selector" dial enables the pilot to select a course along any direction toward or away from the range station, special means being provided to resolve ambiguity. The pilot follows the selected course by keeping the pointer of the indicating instrument centered (vertically).

This facility may be used for determining the true bearing of the aircraft with respect to the station by tuning in the station and adjusting the azimuth selector so that the instrument pointer centers, *i.e.*, reads on course. The selector dial then presents the desired bearing reading. An "azimuth indicator," operating independently of the selector dial and course indicator and giving direct-bearing information, is also available. Bearings obtained from two or more stations will enable the pilot to determine his geographic position.

VOR operates within the band of 112 to 118 Mc. In appearance, power output, and usable distance the VOR station is similar to the VAR station.

The station identification of the VOR type is transmitted intermittently on the carrier, utilizing a 1,020-cps keyed tone, except during voice transmissions. Simultaneous voice transmissions are made on the range frequency of all VOR stations unless otherwise designated by inclusion of the letter W in the station-type designation.

7. Distance-measuring Equipment (DME).* This device, an adaptation of wartime radar, enables the pilot to determine at all times his distance from the transmitting station. The air-borne equipment comprises a transmitter and receiver, the former unit transmitting short high-amplitude pulses approximately 150 times per sec. These pulses are received at the ground station, which likewise comprises a transmitter and receiver, the receiver being tuned to the frequency of the air-borne transmitter. Each time a pulse is received on the ground, the associated transmitter is immediately triggered and sends a pulse reply. When this pulse from the ground is received by the aircraft, circuits in the air-borne equipment measure the time interval between the interrogating pulse originated by the aircraft and the reply received from the "transponder" on the ground. This time is indicated on a meter calibrated in miles.

The air-borne equipment consists of a transmitter, receiver, indicator meter, and associated gear. The ground equipment consists essentially of a receiver and transmitter. Current equipment of this type operates in a band centered at 1,000 Mc. In operation, a beacon is capable of replying to as many as 100 air-borne units at the same time.

The ideal in navigation of aircraft under instrument conditions is to obtain continuous direct readings of distance and direction from one or more ground points. The omnidirectional beacon provides a solution of the problem of direction; the solution of the problem of distance measuring can be achieved by the installation and use of DME. The system provides flight paths especially convenient for accurate manual or automatic following: (1) along radial courses to or from the ground stations; (2) along circular orbits of any radius about the ground stations; and (3), with an R/θ computer added, along any straight-line path offset with respect to the ground stations.

8. Loran. This is a long-range navigation system developed during the war and widely adopted throughout the world. It operates on the following principles: radio

* DME of this type has been decommissioned in accordance with the Air Coordinating Committee's decision of 1956 and will be replaced by a Tacan-type DME, often referred to as DMET. DMET constitutes one component signal of the VORTAC system. As planned, a complete VORTAC installation will consist of co-located VOR and Tacan equipments, thereby providing bearing from either the VOR or Tacan as well as distance (DMET) from Tacan.

DME and DMET utilize the same basic principles, the main difference being that DME is pulse-multiplexed to provide the required number of channels whereas DMET operates with clear-frequency channels.

As of 1958, the CAA has scheduled the implementation of a complete VORTAC environment to cover the airway system of the United States.

signals consisting of accurately synchronized short pulses (80 μsec) are broadcast from a pair of special transmitting stations (a "master" and a "slave"), separated by from 200 to 400 miles and operating on a common frequency within the band 1,800 to 2,000 kc. Each pair of stations comprises a loran facility, or "rate." These signals are received aboard the ship or aircraft on a special receiver, and the difference in times of arrival is measured on a special indicator. This measured time difference is utilized to determine directly, from special loran tables or charts, a line of position which (neglecting the oblate curvature of the earth's surface) is a hyperbola with the two transmitting stations of the loran pair as foci. Two intersecting lines of position, determined from two properly selected loran rates, are required to obtain a loran fix.

A number of loran rates operate on the same radio frequency, but the pulse-recurrence rate differs for each facility. By setting the receiving equipment for the proper frequency and recurrence rate, the navigator is assured of the identification of the facility utilized.

L-f loran is an adaptation of loran at lower frequencies, primarily to increase the range of useful operation. L-f loran has had successful operational service in central and northern Canada on a frequency of 180 kc, and experiments have been conducted on still lower frequencies. One method of pulse matching, called "cycle matching," affords unusual possibilities of accuracy and will be described in detail in a later section.

9. Marker Beacons. Radio-marker-beacon stations of two essentially different classes, v-h-f and l-f, are located along the civil airways to supplement the radio-range system.

V-h-f markers operate on 75 Mc and include the following types:

Station location marker, designation Z. Serves to locate positively the l-f/m-f range station with which it is identified. Nominal power output 5 watts, unkeyed, 3,000 cps modulation, vertical-cone radiation pattern.

Fan marker, designation FM. Serves to furnish aircraft with a check as to progress along airways, to mark junction of two radio-range courses, to designate "hold-over" point along airport approach 10 to 20 miles from airport, or to mark obstructions. Nominal power output 100 watts, keyed for identification, 3,000 cps modulation, fan-shaped radiation pattern extending (at 1,000 ft altitude) 2 to 4 miles along the range course and 12 miles across the course.

Low-power versions of the fan marker, designation LFM. Nominal power output 5 watts or less; used for special purposes such as for "let-down" along a radio-range course. The lower power is required for installations closer than 8 miles to an airport to reduce interference with Z marker.

L-f markers are nondirectional radio beacons operating in the band 200 to 400 kc and serving principally for use with automatic direction finders (ADF) aboard the aircraft. Station identification, utilizing a 1,020-cps keyed tone, is transmitted intermittently on the carrier. The several types are:

Type HH. 2,000 watts or greater.

Type H. Power output 50 watts or greater; voice-transmission feature unless designation is followed by letter W; operates continuously.

Type MH. Power output less than 50 watts; operates on request, usable range 25 miles.

10. Instrument-landing System (ILS). This was developed to enable properly equipped aircraft to navigate to a safe landing during poor weather conditions without visual reference to the ground. It serves also to reduce the length of time aircraft must be held aloft, during periods of congestion, by speeding up instrument approaches and landings.

A standard instrument-landing system ordinarily consists of a localizer unit (LO), a glide-path unit (GP), and two position markers, each with a homing beacon, designated as "outer marker" (OM), and "middle marker" (MM). In some cases, the homing beacon at the middle marker is omitted.

The *localizer unit* provides a course for horizontal guidance. It is similar to a two-

course v-h-f radio range aligned with the direction of approach. It operates in the band 108.3 to 110.3 Mc, with a maximum power output of 200 watts, and is of the "equisignal" type. It actuates the vertical pointer of the cross-pointer instrument on the instrument panel of the aircraft, thereby providing continuous indication of the lateral position of the aircraft with respect to the approach course. A simultaneous voice feature, for use by the airport controller, is provided on the localizer, together with station-identification keying when voice transmissions are not being made.

The *glide-path unit* provides an equisignal course in the vertical plane containing the approach course and inclined at an angle of 2 to 4 deg with the horizontal as measured from the point of contact. It actuates the horizontal pointer on the aircraft cross pointer and provides continuous indication of the vertical position of the aircraft with respect to the desired "let-down" course. It operates in the band 332.6 to 335.0 Mc with a maximum power output of 20 watts.

Marker beacons are of the low-power v-h-f fan-marker types operating on 75 Mc, with power output of 2 watts. They are tone-modulated and identification-keyed, and may be identified aurally or by lights on the instrument panel of the aircraft. The fan-shaped pattern is directed vertically upward with the fan at right angles to the approach course. The outer-marker modulation frequency is 400 cps and its identification signal consists of 2 dashes per second. The middle-marker modulation frequency is 1,300 cps and it is keyed continuously with alternate dots and dashes. The two markers are located along the approach course, as follows: outer marker, 4 to 7 miles from the approach end of the runway and within 150 ft of the runway center line extended; middle marker, 3,500 ft (±250 ft) from the approach end and within 50 ft of the runway center line.

Homing beacons (type K) are nondirectional transmitting stations in the band 200 to 400 kc, with power output of 25 watts, and operate continuously. They provide signals for automatic radio direction finders aboard the aircraft and thus furnish means for determining the heading of the aircraft with respect to the desired approach course.

Some 157 ILS installations were in operation in the United States as of January, 1956.

11. Ground-controlled-approach System (GCA). Developed by the armed services during the war, some 31 (PAR and ASR) CAA GCA installations were available to civil aircraft as of January, 1956, either on a continuous or on an emergency (30-min notice) basis. The GCA system is a specially developed combination of search and precision scan radar units which permit its operators to "talk the airplane down" to the point of contact on the airport surface. Only the regular flight instruments and communications equipment are needed on the aircraft.

In the wartime version, the GCA station is placed 500 ft to the side of the runway. A 10-cm search radar scans a 35-mile area around the airport, and its operator directs pilots of inbound aircraft to the approach area at correct altitude. This area is about 6 miles from the approach end of the runway.

Two precision 3-cm radar beams scan the approach sector in azimuth and elevation. These give the ground operators precision measurements of distance and lateral and vertical displacement of the aircraft. By proper voice instructions to the pilot, the aircraft is controlled by the ground operators to follow an approach path corresponding to a line drawn on the radar scope maps.

More recent versions of GCA, improved to conform more closely to civil-aviation requirements, employ a three-dimensional scope with the information remotely indicated in the airport control tower. This eliminates the need for coordination between a number of operators and provides for control of all traffic by the traffic-control tower.

Because civil aircraft may not be able to receive the voice transmissions on Army and Navy assigned frequencies, an emergency v-h-f transmitting channel has been provided on 140.58 Mc. The pilot desiring GCA service contacts the CAA airport traffic-control tower. The GCA voice transmissions are then made on the regular assigned station transmitting frequencies. For aircraft not equipped to receive either h.f or v.h.f, the instructions may be relayed on CAA facilities in the 200- to 400-kc band.

12. Direction Finders. Radio direction-finding facilities (DF) represent auxiliary aids to civil aviation available in two forms: the direction finder may be on the aircraft or it may be on the ground. For requisite accuracy with simple antenna equipment aboard the aircraft, direct space-wave reception is required; accordingly, low radio frequencies or very high radio frequencies must be employed. Since ground DF installations may have receiving antennas of any required degree of complexity, there are no restrictions as to frequency bands. High frequencies in the range 2 to 30 Mc may be utilized with sky-wave distance ranges of hundreds, or even thousands, of miles. Such distance ranges are particularly useful for transoceanic navigation. However, for overland operation, expediency has indicated the use of the direct space wave with corresponding ranges of the order of 100 miles.

Air-borne DF is utilized for homing purposes and for cross-bearing fixes. A four-band set, to cover the frequency range 100 to 1,759 kc, provides automatic bearing indication of the direction of arrival of r-f energy and simultaneous reception of intelligence. Such a unit may be utilized on any of the CAA 200- to 400-kc facilities with range depending upon the power of the transmitting station, with the exception that the loop ranges (RL, MRL, ML) are not suitable for direction-finder operations.

Another important facility that can be utilized by aircraft pilots for purposes of direction finding and fix determination consists of regular transmissions from commercial broadcast stations. All such stations having a power rating of 100 watts or more are listed in the "Flight Information Manual," both by operating frequency and by geographical position. Satisfactory use of broadcast-station transmissions as navigational aids is dependent upon a proper understanding of the properties and peculiarities of radio direction finders, inasmuch as reflected sky waves often produce serious interference with ground-wave signals.

Ground-station DF takes bearings on transmissions from the aircraft and, by triangulation between two or more stations, provides the aircraft with a position fix.

13. Organization of System Descriptions. Electronic avigation aids fall naturally into four categories which represent certain phases of flight. Equipment descriptions will accordingly be presented under the following functional classification: Part 1, The Equipment of the En-route Long-distance Zone. Part 2, The Equipment of the En-route Short-distance Zone. Part 3, The Equipment of the Approach and Landing Zone. Part 4, The Equipment of the Airport Zone (Surface Detection).

Very large expenditures have already been made by governments and by the aviation industry in establishing extensive communication systems, flight-path beacons, landing aids, and schemes for traffic control. Responsible agencies must consider these present facilities from the point of view of possible continued utilization and expansion, rather than from abandonment in favor of newer systems. This fact has been given wide recognition in the long-term planning which is being conducted by the Special Radio Technical Division of the International Civil Aviation Organization (ICAO). This division, on which representation is world-wide, has made an extensive study of civil air-navigation requirements along international air routes and has prepared specific recommendations covering the evolution of a broad program of expansion and development. The findings of its various committees represent the consensus of authorities on aircraft navigation and serve as a background for much of the material presented here.

PART 1

THE EQUIPMENT OF THE EN-ROUTE LONG-DISTANCE ZONE

The en-route long-distance zone has been defined[1]* as that portion of the airspace in which aircraft are primarily served by radio aids spaced at intervals of 200 miles or more, exclusive of the en-route short-distance, approach, and airport zones. A promi-

* Numbered references appear at the end of this chapter.

nent example of an en-route long-distance zone is a transoceanic area where the installation and maintenance of navigation facilities at short intervals would be impractical. Long-distance zones also include many land areas throughout the world where it is uneconomical to locate navigational aids at intervals less than 200 miles and indeed such installations may not be justified where traffic densities are low.

The requirements of long-distance navigation aids are not limited to directing the aircraft to an intended destination. For optimum efficiency, the system must be capable of directing the aircraft along minimum flight paths[2] so as to conserve fuel and time in flight. This indicates not only the ability to define a great-circle route but flexibility to deviate from this planned route so as to take advantage of pressure systems and wind currents aloft as may be reported by weather forecasts and observations.

In most cases, the navigation system itself must make possible the determination of the pressure (wind direction and velocity) pattern. This determination is normally based on two or three successive fixes and is therefore dependent on the relative fix accuracy rather than on the absolute fix accuracy. Therefore, it is important that the navigational system should give good relative accuracy between fixes (to determine ground speed and direction) as well as absolute geographical fix accuracy.

In establishing requirements for a long-distance system of air navigation the Special Communications Technical Division of the International Civil Aviation Organization (ICAO)[3] stated that the system shall

1. Make continuously available an indication of the aircraft's geographical position by a recognized method of fixing relative to the ground
2. Provide to the pilot a continuous visual indication which will enable him to follow any operationally desirable track
3. Be capable of providing for either automatic or manual flight
4. Be capable of satisfactory operation despite atmospheric or propagational disturbances and the effects of terrain
5. Be capable of being integrated into an over-all system providing automatic take-off, en-route flight, position reporting, air-traffic control, and landing
6. Provide coverage with visual or aural indications to a distance of at least 1,500 nautical miles over water and 750 nautical miles over land
7. Have an order of accuracy such that position-fixing error will not exceed 10 nautical miles at distances greater than 1,000 nautical miles, or 1 per cent of the distance between 1,000 nautical miles and 500 nautical miles, or 5 nautical miles at distances less than 500 nautical miles

In 1954, the ICAO Communications Division at its fifth session[4] recommended that the operational objectives listed above be reviewed, particularly in view of the probable performance of the systems then under development. In the material that follows, all the systems in use throughout the world, together with some of the newer developments, are described.

14. Air-borne Direction Finders and Radiophares. It is common practice in the United States to use the l-f four-course radio-range station as a source of radiation on which bearings can be taken with aircraft direction finders, but in other parts of the world special stations are often erected for this purpose. These stations operate on frequencies from 200 to 1,600 kc.

Nondirectional ground stations used as a source of radio transmissions on which bearings may be taken are referred to as *radiophares* and are probably the simplest form of navigational aid. Stations of this character form the backbone of navigational systems in South America and in various parts of the globe. In the United States, plans have been discussed for the installation of high-powered nondirectional stations as auxiliary aids when the four-course low-frequency radio ranges are decommissioned.

15. Loop Receiving Antenna. The loop antenna is the device that forms the basis of all low-frequency aircraft direction finders. Reference to Fig. 2 indicates that the receiving loop antenna is a winding composed of any number of turns of wire. Since an incident radio wave contains a magnetic-field component, the loop can be considered as a secondary of a transformer subjected to a magnetic flux of varying

intensity. The magnetic field in a normally polarized wave is at right angles to the direction of propagation. If the wave is traveling parallel to the plane of the loop ($\theta = 0°$) it will cut first one side and then the other and produce the greatest phase difference in the voltages induced. If, however, the plane of the loop is at right angles to the direction of wave travel ($\theta = 90°$) the magnetic field will cut both sides of the loop simultaneously and generate the maximum voltage.

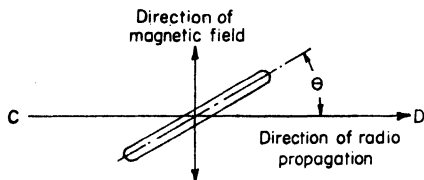



Fig. 2. Relation of loop antenna to direction of wave propagation.

If the voltage is plotted as a function of θ, the resulting curve will be the familiar figure-of-eight pattern. That is, as a loop receiving antenna is rotated about its vertical axis, there will be two positions for which the output is 0 and two for which it is maximum.

It can be shown that, whether the induced voltage is considered a phenomenon caused by the electric or magnetic field, the results are the same. In essence, detecting the voltage induced in the loop, and determining its variation with loop orientation, forms the heart of a direction-finding system.

In air-borne radio-compass work, a shielded loop is used to improve the electrical properties and to serve as an ideal mechanical construction capable of withstanding aircraft service. The metallic cover forming the shield has a break at one point and is connected to ground. The circulating currents induced in the shield are negligible for all practical purposes. The magnetic field penetrates the shield and acts upon the conductors within to induce a voltage as discussed above. Hence, the figure-of-eight pattern will continue to prevail.

16. Simple Direction Finder. In its simplest form, the aircraft direction finder consists of a loop antenna (usually of the shielded type), a loop bearing indicator, a radio receiver, a receiver tuning unit, and a loop rotating gearbox. In operation the receiver is tuned to any desired station and the loop is rotated until the signal disappears. The bearing obtained is then compared with a magnetic compass so that its value as a function of north may be known. Two such bearings obtained from suitably located stations determine the location of the aircraft and constitute a "fix."

The bearing is obtained from the null rather than the maximum of the signal since the null has greater sensitivity with respect to azimuth and provides better accuracy.

17. The Radio Compass. The basic elements of the radio compass were first described by G. G. Kruesi[5] in a patent disclosure of November, 1930. In essence, the system depends upon the difference of the phase of induced voltages in a loop antenna as opposed to an open antenna. Because of its electrical characteristics, a loop antenna provides a voltage 90 deg out of phase with the voltage induced in a nearby open antenna. When combined in the proper proportion this receiving system provides a cardioid pattern and thereby resolves the bidirectional ambiguity.

In practice, the loop is mounted so that its plane is at right angles to the major axis of the aircraft. Both the loop and the vertical antenna are connected to a special radio receiver. The elements of operation are indicated in Fig. 3. If the loop is connected to the receiver by means of a reversing switch, the field pattern of the antenna system (the cardioid) can be made to reverse. If a meter having two opposing windings is connected to the rectified output of the radio receiver, and the connections to this meter are changed as the loop reversing switch is changed, the meter will move to the right or left depending upon the position of the transmitting station. If the nose or tail of the aircraft is pointing to the radio station the indicator will read zero. If the reversing switch is moved with sufficient rapidity, it will deflect only an amount proportional to the difference between the receiver outputs for the two loop connections. In actual equipments the reversing mechanism is normally done by electronic means. In practice, the sensitivity of this system is adjusted so that a deviation of 5 deg from directly ahead produces full-scale deflection of the instrument.

Practical equipments use a system of modulating the r-f energy as shown in Fig. 4. With this circuit the deflection of the indicating meter is determined by both the amplitude and phase of the audio tone. For a station dead ahead the audio tone output is zero. This tone may also be connected to headphones and thereby provide the

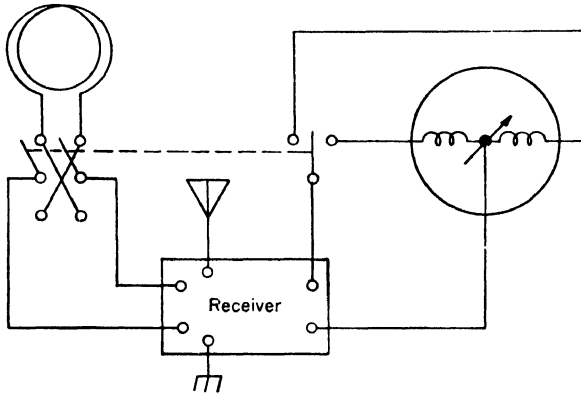



Fig. 3. Principle of the right-left compass.

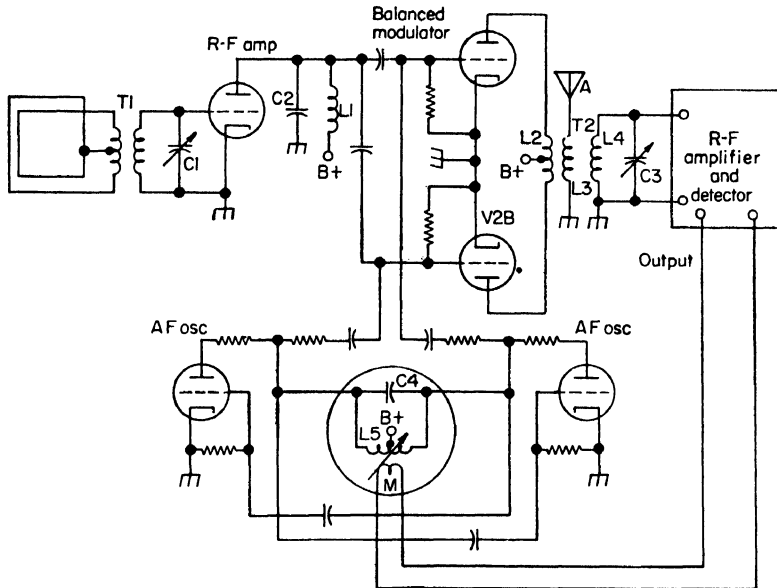



Fig. 4. Elements of the right-left compass.

pilot with an aural output which is zero only when the aircraft is pointing directly at the station.

18. Automatic Direction Finder. The right-left radio compass was never widely adopted by commercial-air-transport operators or by military operators. Its major disadvantage is that, in the presence of a cross wind, the aircraft traverses a path much longer than the shortest path between the starting point and destination. This homing procedure can be minimized by continually referring to a magnetic compass, but this procedure detracts from the usefulness of the radio compass. The automatic

direction finder was suggested by F. L. Moseley[6] on Oct. 22, 1937. Several devices of this type have been developed and have found wide acceptance by the air lines. The principle of operation is shown in Fig. 5. In essence, it consists of a sensing device such as a contact on either side of the moving needle. As the needle touches one of the contacts, by virtue of the fact that the radio station is to one side or the other, a motor is set in motion. This motor rotates the loop which, unlike that of the right-left radio compass, is not fixed. Hence, the main difference is the movement of the loop rather than the reorientation of the entire aircraft. When the loop is rotated until its plane is at right angles to the direction of propagation the needle returns to zero and the motor stops. By displaying the position of the loop, the bearing between the heading of the aircraft and the radio station is indicated. By two or more such readings, the aircraft position may be fixed.

FIG. 5. Principle of the automatic direction finder.

Actually the circuit of Fig. 5 merely serves to illustrate the principle. Electronic circuits are substituted for the contact mechanism described above.

19. Radiophares. The ground station used with aircraft direction finders is termed a radiophare. A radiophare must produce sufficient field strength to permit the satisfactory operation of the air-borne direction finder under the conditions of atmospheric noise normally encountered. With narrow filters provided in automatic direction finders, good operation can be obtained with a signal-to-average-atmospheric-noise ratio of the order of 1:1, provided that the field strength is not less than 20 μv per m. During daytime, the noise will normally be less than 20 μv per m. Hence, the limiting condition is the inherent noise of the receiver. However, beginning just before sunset, the atmospheric noise level rises sharply and reaches a value of approximately 100 μv per m. Therefore, it is necessary during the night hours for the radiophare to produce a field strength of at least 100 μv per m to ensure satisfactory operation.

As an indication of the range capability of radiophares the following approximate figures are given: A 1-kw transmitter operating at 300 kc will normally produce 100 μv per m over mountainous terrain to a distance of 100 miles, over fertile soil to more than 500 miles, and over sea water to greater than 1,000 miles.

20. Four-course Low-frequency Radio Range and Markers. The four-course l-f radio range has been the primary device for furnishing guidance to aircraft in the United States since 1929. These stations operate in the band from 200 to 415 kc, where only approximately 65 channels are available. For this reason,[7] in the United States it was necessary to reduce power so that mutual interference would not result. Because of this power reduction, the l-f radio range is sometimes considered to be an en-route short-distance aid. However, with sufficient power it is capable of giving guidance over long distances.[8]

Because the four-course radio range is essentially a homing device, some alternate means must be employed to obtain a fix while flying the facility. With ranges alone, a fix can be obtained only over the station (termed the *cone of silence*) or at intersections of ranges. As the airway system expanded, fixes became necessary at other than these limited points. Markers were developed for this purpose which radiate a cone of signal rather than one of silence. These markers can be used to mark various points

along the courses of the l-f range. While it is true that markers have other purposes than described above, they are normally closely associated with the operation of a four-course radio range and are therefore described in this section.

21. Radio Range. Early radio ranges utilized two loop antennas spaced at right angles to each other. The United States ranges employed polygonal loops 50 ft high,

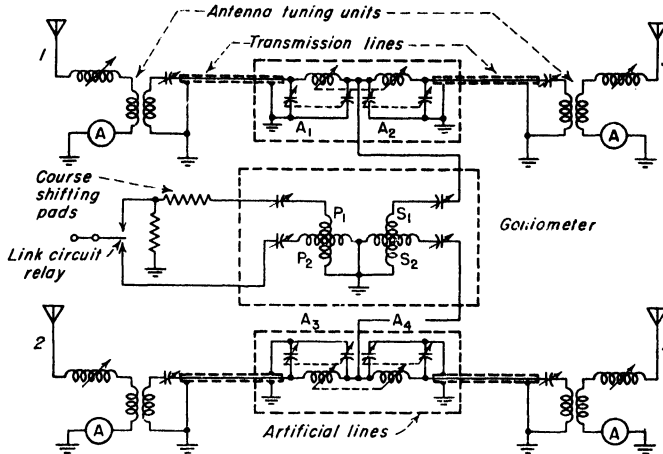



FIG. 6. Range-beacon antenna-coupling system.

200 ft across the base, and 30 ft on the extreme sides. Each loop produced a directive pattern referred to the plane of the loop. Although the two patterns were identical, they were displaced in space by 90 deg because of the loop orientations. Therefore, at 45 deg to each loop, and only at this point, signals received from both loops would be equal. The early equipment made use of this principle by utilizing a modulated transmitter which was alternately connected first to one loop and then to the other. If the dwell time were different, for example, 3 sec on the first loop and 1 sec on the second loop, a Morse code resembling A or N would be audibly received to either side of the course. For the 45-deg position the switching of the transmitter from loop to loop would not be observed and a continuous tone would be heard.

To minimize night effect, the loop antenna was replaced by vertical steel towers fed by buried transmission lines. The complete antenna system comprises five self-supporting base-insulated steel towers approximately 130 ft high, four of which are placed at the corners of a square 300 to 500 ft on a side, and the fifth at the center of the square. The essential portions of the antenna system are shown in Fig. 6. The principle of operation of the radio-range beacon is evident from Fig. 7. The four outside

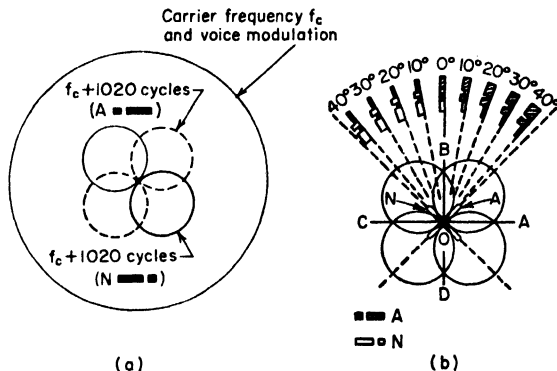



FIG. 7. (a) Space radiation pattern of simultaneous l-f directional range-beacon and voice broadcast station. (b) Range-beacon signals produced.

antennas form overlapping figures of eight when operated in pairs of two. The intensities of the side-band emissions (formed by the two directional-antenna systems), and hence the intensities of the detected signal produced by beating the side-band emissions with the nondirectional carrier, are equal along the lines which bisect the angles between the two antenna systems. An aircraft may follow these bisectors if means are provided for distinguishing the radiations from the two directional antennas. For this purpose, an automatic keying relay keys the radio power to one of the directional antennas in accordance with the Morse character N and the power to the second directional antenna in accordance with the Morse character A. The coded signals are sent out in groups and are interlocked so that along any of the four courses they form a continuous signal interrupted every 35 sec by the station-identification signal. The "on-course" signals are obtained along zones 2 to 3 deg wide. Off the course, the monotone signal breaks up into the component A and N signals, one or the other being of greater intensity depending upon the side "off course."

The coupling system between the transmitter and the directional antennas incorporates a keying relay, a goniometer, a course-shifting pad, artificial line sections, concentric transmission lines to the tower antennas, and antenna-coupling tuning equipment. The relay is energized by a motor-driven keying device so as to key the r-f power to the primary windings P_1 and P_2 of the goniometer in accordance with the required N-A sequence indicated in the foregoing description. The goniometer offers a convenient method of orienting the beacon space pattern, and the course-shifting pads and artificial line sections are used for shifting the range-beacon courses from their normal 90-deg relationship in order that they may be aligned with the required airways. This course shifting permits the angle to be placed between 60 and 120 deg rather than being fixed at the normal 90-deg position. In this way, the legs of the range can be arranged so as to bear on airways leading from or to a city. Precautions, in the form of special automatic phase and amplitude control of the antenna currents, are provided to ensure stability of the space pattern.

The transmitter employs two independent r-f channels controlled by two matched quartz crystals spaced 1,020 cps apart. Continuity of service is ensured by means of complete standby equipment provided with an automatic transfer relay. In the absence of speech modulation, the setup forms a single-side-band system having 1,020-cycle modulation. The carrier is radiated nondirectionally by the center tower, whereas the side band has the characteristic radiation of the radio range beacon. When speech modulation is applied, the central tower radiates these side bands nondirectionally. Therefore, the system provides simultaneous weather broadcast and directional guidance. Appropriate filters are used both in the ground transmission and in the air-borne receiver to eliminate interference between the speech frequencies and the 1,020-cycle beacon signals.

22. Marker Beacons. To correct the difficulties associated with the identification of the cone of silence over a range station, there was installed and commissioned on Jan. 1, 1939, on the airways of the United States a different type of terminal indication known as the *Z marker*. The marker operated at 75 Mc and consisted of a transmitter on the ground which actuated a separate receiver on the aircraft. Primarily the indication in the cockpit consists of a light that illuminates as the aircraft passes over the transmitting array, but an aural indication in the form of a continuous 3,000-cycle tone is simultaneously heard. Dual transmitting equipments are employed and have a nominal output of 5 watts. Automatic transfer is arranged in the event of a failure of one transmitter.

The antenna array consists of four horizontal half-wave elements connected so that one pair differs 90 deg in phase from the other pair. In this manner essentially circular polarization results. The array is mounted a quarter wavelength above a wire screen that serves to project energy upward.

The demand for fixes while flying the radio range, at points other than the terminal, was first met by use of intersections produced by two range systems. Special l-f radio ranges were employed for a time solely for marking purposes, but the lack of sufficient l-f channels was a big disadvantage. The problem of unlimited fixing facilities was met by the use of a 75-Mc marker not unlike the Z marker previously described except

that this so-called *fan marker* was placed on an airway instead of at the radio-range position.

Because an aircraft might be to the right or left of the "on course" of the radio range, it is necessary that the marker for these range legs have some breadth extending at right angles to the major dimension of the range course. At a 1,000-ft altitude, the radiation pattern extends 2 to 4 miles along the range course and approximately 12 miles across the course. In order that the fan indication might be received over the required area without the necessity for changing receiver sensitivity, it was essential that the power of its transmitter be greater than that of the Z marker. Consequently, the fan-marker transmitter is rated at 100 watts.

The antenna forming the fan pattern consists of four half-wave elements all located coplanar and with their longest dimensions in line. Each element is excited in phase with the element next to it. The entire array is located a quarter wavelength above a galvanized-iron screen.

In addition to furnishing aircraft with a check as to progress along airways, the fan marker is also used to mark the junction of two radio-range courses, to designate hold-over points along an airport approach, or to mark an obstruction.

A low-power version of the fan marker, with a nominal power output of 5 watts or less, is used for special purposes such as let-down along a radio-range course or localizer course.

23. Consol (Sonne). Consol represents one of the more important l-f systems of navigation in the long-distance zone. The system was developed by Germany early in the Second World War and was used extensively under the name of Sonne. The English radio observers deduced the significance of these radiations and made use of the signals for navigating aircraft of their own forces. For security reasons, the system was renamed Consol. The greatest advantage of the Consol system is that it is capable of giving accurate bearings over long distances and yet the only air-borne equipment required is a standard communications-type receiver. Its greatest disadvantage is that a comparatively long time (in excess of 30 sec) is required to obtain a reading of bearing and at least double that time is required to obtain a fix by establishing bearings to two stations. The system does not produce a visual presentation of the bearing data and cannot readily be adopted for automatic guidance of aircraft.

As late as 1954, Consol was one of the systems considered for international standardization by the ICAO.[9] However, it was not standardized, but was approved for long-range use together with loran and radiophares.

The basic antenna system comprises three radiators, as shown in Fig. 8, spaced 3λ apart. The antenna currents in the outside radiators are one-fourth of the current in the central radiator and 90 deg leading and lagging in phase, respectively. This antenna combination produces the pattern shown in Fig. 9.

In the initial system (called Elektra) the phases of the currents in the outside radiators were successively reversed according to a dot-dash time cycle so that one of the overlapping patterns was characterized by a series of dots and the other by a series of dashes. At the intersections, the dots and dashes merged into a continuous signal and thereby resembled the operation of the l-f four-course radio range.

The system later evolved into Sonne, in which the phases of the currents of the outside radiators were shifted from their respective 90 deg leading and lagging values continuously in opposite directions during a 1-min interval at a rate of 3 deg per sec. Simultaneously, the currents in the radiators were reversed according to the dot-dash time cycle. There is thus produced a slow rotation of the equisignal lines so that, at the end of the 1-min interval, the dot-and-dash patterns have become completely interchanged and each equisignal line then occupies the position of the one adjacent to it at the beginning of the interval.

In the Consol system, Fig. 9 indicates that coverage is provided for only approximately 240 deg of the total 360 deg. The Consol lobes each occupy a sector of about 10 deg. Every other lobe transmits a different Morse character (E or T), the E signal being transmitted in one-third the time of the T signal. Between the lobes, the interlocking signals produce a continuous tone. With the large number of similar lobes that are present, it is necessary to determine position with an accuracy of at least ±20 deg by using an air-borne direction finder before a more accurate position can be determined. The accurate position

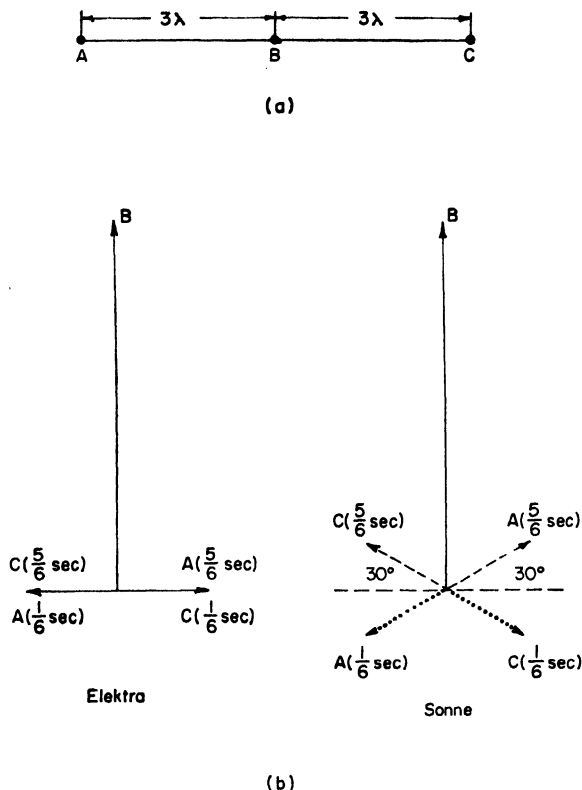



FIG. 8. Antenna spacing and phasing for Consol beacons.

is determined by counting the relative number of E and T characters that precede and follow the "on-course" signal.

The usual cycle of operation of a Consol station is about as follows:

1. Omnidirectional transmission with identifying station letters for 28 sec.
2. Break for 1.5 sec.
3. Transmission of 30 each of the E and T characters while the lobes rotate for 30 sec.
4. Break for 1.5 sec.

Specially prepared charts are available to convert E or T total counts into bearing for each station.

The transmitting equipment used for generating the Consol signals is essentially a standard nonmodulated c-w transmitter with crystal control operating in the 200- to 300-kc band. The output of the transmitter is connected by a transmission line to the tuning units of the central radiator and also to the Consol signal-generating unit. The signal-generating unit in turn is connected to the two outer radiators. Maximum output of the usual transmitter is about 2 kw. However, some equipments having a power output of up to 50 kw have been planned. A presentation of the apparatus arrangement is shown in Fig. 10.

With the use of simple narrow-band airborne receivers having sensitivities of the order of 1 μv per m, the Consol system provides services to approximately 400 miles over poor terrain, 800 miles over good terrain, and 1,000 miles over the sea during daytime. At night, sky-wave propagation becomes a factor and ranges in excess of 1,000 miles can be realized for all types of terrain. Under normal conditions, and excluding propagation anomalies that occasionally take place, bearing accuracies within 1 to 2 deg are possible.

FIG. 9. Field pattern of Consol station.

Many descriptions of improved forms of Consol have appeared in technical publications. One such system, developed in the United States, was installed largely to

obtain propagational data but has not been employed for use in routine air operations. It consists of a twin two-tower system with one tower doing duty in both systems. All three towers are in line. Normally only the dual-purpose tower and the outermost radiators are energized. The spacing between these towers is three wavelengths and the pattern produced in space is identical to that described. At intervals, however, the intermediate tower (located one wavelength from the dual-purpose radiator) is energized in place of the outermost radiator. With the closer-spaced towers, a four-lobed pattern is produced that may be employed to resolve ambiguity without the use of a direction finder.

The United States system also employs several other novel features. A goniometer is connected at the output of an oscillator instead of to the output of the transmitter. Transmission lines connect the oscillator to 10-kw power amplifiers located at the base of each of the three radiators. In this manner the problem of handling high powers

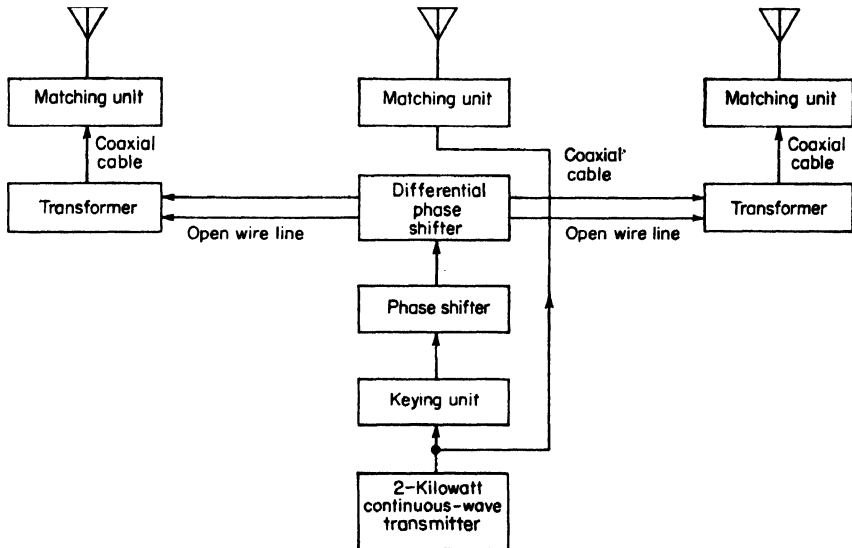



FIG. 10. Consol transmission system.

with the goniometer is avoided. By use of this technique, error produced by transmission-line radiation is also avoided and there are no losses to affect the system efficiency.

24. Some Low-frequency Developments. The existing long-distance navigation aids described previously do not fully meet the requirements of ICAO. Accordingly, a number of devices are in various stages of development which their sponsors hope will provide better solutions and be considered for adoption by ICAO. This section describes the most prominent of these developments which have progressed to the testing stage.

In connection with these developments, studies[10,11] have been made to determine what frequencies are best suited for the operation of long-distance ground-based aids. The fundamental requirement is that reliable and undistorted signals must always be present, regardless of propagation conditions, to a distance of at least 1,500 miles from the ground station. Frequencies above 30 Mc are unsuitable because of typical line-of-sight limitations. The daytime attenuation of frequencies between 400 and 1,500 kc is also considered unsuitable because of the excessive power required. Frequencies between 1,500 and 30,000 kc are attractive because of high-efficiency antenna structures, low power, and low static levels. However, these frequencies are charac-

terized by skips wherein the signal is not efficiently reflected back to earth at certain points, and simultaneous transmission[12,13] on several frequencies would be required. Also, these frequencies have been found to be susceptible to rapid fading.[14,15] Studies have concluded, from data accumulated over a period of years, that frequencies between 10 and 150 kc offer the most reliable service, except for certain climatic and seasonal variations.[16] Taking into account economic factors, power requirements, and noise levels, a generally accepted optimum compromise is at frequencies of about 100 kc.

25. CAA Low-frequency Omnidirectional Range. Following the successful development of the v-h-f omnidirectional range, CAA began the development[17] of an l-f device of a similar character. The device was intended for use in the 200- to 400-kc band, when that band was made available by the discontinuance of the l-f four-course range. Unlike the four-course range, the omnidirectional range could supply navigational data at all azimuths, rather than being restricted to four specific courses. The intent was to supply transoceanic and perhaps some national long-distance navigational service.

The principle of operation involves two radiated signals. One of these, transmitted omnidirectionally and having a characteristic modulation, is known as the reference signal. The second signal employs a figure-of-eight field pattern which is not modulated. The figure-of-eight pattern is, however, continuously rotated at a rate of 1,800 rpm, thereby producing a 30-cps modulation at the observing point. The modulation of the first omnidirectional transmission comprises a reference signal against which the phase of the variable or rotating signal can be compared. If the equipment is adjusted so that the voltage of the modulating frequency of the reference signal is a maximum as the figure-of-eight maximum is passing through north, the phase between the reference signal and the variable signal will be directly proportional to the deviation from north.

To avoid the necessity for two carriers, one radio frequency is used and modulated by a 210-cps subcarrier. Accordingly, at a spacing of 210 cps from the carrier, a separate sideband signal appears that can be separated in the radio receiver. The subcarrier is in turn frequency-modulated by a 30-cps signal which, when properly demodulated, provides the reference signal.

The variable-phase field is produced by an unmodulated carrier fed to the rotor coil of an inductive-type goniometer. The two stators of the goniometer (disposed at right angles) feed two pairs of diagonally disposed vertical towers identical to those described for the simultaneous radio range. When the rotor is rotated at 900 rpm, and the two pairs of towers are driven 90 deg out of phase, the resultant radiation pattern is a figure of eight rotating at a 30-cps rate. Since it is imperative that the phase of the reference and the variable signal be in absolute synchronism, the reference-signal generator is driven by the same motor that rotates the goniometer. The reference generator takes the form of a seven-tooth tone wheel which normally produces a frequency of 210 cps. By varying the angular spacing of the teeth, the 210 cps is given a frequency modulation of 30 cps.

The transmitter used with this development was derived from the transmitter of the four-course simultaneous radio range, with both channels operated on the same frequency. The tuning of the antenna array involves the following phase adjustments:

1. The currents in the diagonally opposite towers of a pair must be related by 180 deg and must be equal in amplitude.
2. The phase relation of the currents in the tower pairs must be zero or 180 deg.
3. The phase of the current in the central tower must be 90 deg with respect to the currents in the outer towers.

When the radiated signals are properly demodulated and processed, the air-borne receiving equipment yields two 30-cycle components whose phase difference indicates the bearing from north. A bearing selector is used to provide a means for indicating the bearing. This comprises a phase shifter which changes the phase of the reference signal as received, until it has the same phase as the variable signal. To make this measurement sensitive, a zero-phase detector is used consisting of a phase-discrimination circuit and a zero-center d-c meter. By rotating the bearing selector so that the meter indicates zero, the bearing may be read. Conversely, a desired homing track may be set into the bearing selector and the aircraft flown so that the indicator remains centered.

The system was intended to provide a useful range of 600 miles or greater overland with an accuracy of within ± 3 deg. Tests of the quadrantal errors made at a distance of 6 miles

from the station showed them to be of the order of ±2 deg. Recordings made at a distance of up to 570 miles failed to disclose selective fading between the two components of the signal.

26. Post Office Position Indicator (POPI). Early in 1941, the British Post Office was asked by the Admirality to consider the development of a precision navigational aid for shipboard use under conditions of low visibility. The result of this work was a c-w phase-comparison system known as POPI.[18]

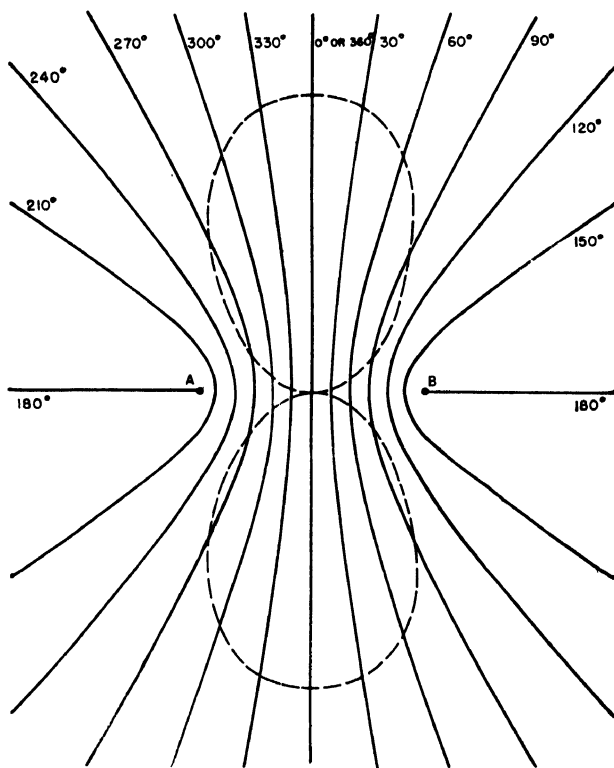



FIG. 11. Lines of constant phase difference and radiation pattern in horizontal plane for radiators spaced $\lambda/2$ apart and in phase synchronism.

System operation is based on the hyperbolic principle which is illustrated in Fig. 11. If two radiators A and B, spaced $\lambda/2$ apart, are excited in phase from the same transmitter the phase of the signals will remain zero along the perpendicular bisector of the array. Conversely, along the line of the array, the phase difference will be 180 deg. Between these two positions, all phases between 0 and 180 deg exist, and the loci of all equiphase points form a family of hyperbolas. Accordingly, by knowing the phase difference and wavelength, it would be possible to determine the hyperbolic line of position on which the observer is located with respect to the transmitting station.

Since the hyperbolic lines are not equally spaced for a two-antenna system, it is apparent that accurate bearing indications cannot be obtained at all azimuths. To correct this condition, the POPI system utilizes radiation from three antennas arranged at the corners of an equilateral triangle. The sequential transmission from pairs of the three-antenna system then supplies nearly equal accuracy at all azimuths. In addition, by making readings on a second pair, the ambiguity that is present can be eliminated.

Since it is not practical to measure the phase difference between two transmissions on the

same frequency, POPI resolves the problem by alternately keying each radiator with dashes, followed by a pause. The pause serves to identify the dashes; that is, the first dash after the pause is known to radiate from A, the second from B, etc. The system then transmits a second carrier frequency, at a constant frequency difference from the first, and these two frequencies when beat in the receiver form the desired audio frequency for measurement of phase difference. The carrier frequencies differ by a value of 83⅓ cps maintained constant by a monitor and control circuit referenced to a tuning fork.

In the receiving system, the 83⅓-cps tone provides a means for separating the transmitted dashes. The tone drives a synchronous motor which, in turn, commutates and distributes the outputs of the receiver to the phase-measuring circuits. Although this provides the proper speed of rotation, a proper starting phase is also necessary and is accomplished by reference to the blank position in the sending sequence.

To determine bearing, a means is necessary for reading the phase difference between the radiated signals although they are not present simultaneously. In the transmission, this is provided for by a commutator rotating at 5⅚ rpm driven from the 83⅓-cps frequency. The commutator then keys the oscillator frequency to the amplifiers located at the various radiators. The result of this modulation is a wave consisting of a frequency of 83⅓ cps with side bands of 5⅚ cps. By using a narrow band-pass filter to separate out the low-frequency side bands, the carrier frequency of 83⅓ cps is restored in the air-borne equipment. If the length of each dash is made an integral number of half cycles of 83⅓ cps, the phase of this frequency will be the same as that of the original r-f carrier. With the transmissions separated by the commutator, these signals may be connected to a standard phase meter to indicate the phase difference directly, or a null meter similar to that described for the CAA omnirange.

The only data available to date are a series of measurements made relatively close to the transmitting stations. These measurements taken at all azimuths show most errors to be under 1 deg but some, because of site errors, were as much as 15 deg.

27. Navaglobe-Navarho. These are the names applied to a long-distance l-f aid developed by the Federal Telecommunications Laboratories.[19,20] Navaglobe applies to the system when bearing alone is furnished, but if an additional translator is added to the air-borne equipment it becomes a Navarho system which can supply both bearing and distance. Official recognition was given to this system by the fifth session of the Communications Division of ICAO, which met in 1954.[21] Although not adopted as a standard, it was listed as a system that should be developed further to qualify it as a candidate for the ultimate common-standard long-distance navigation system.

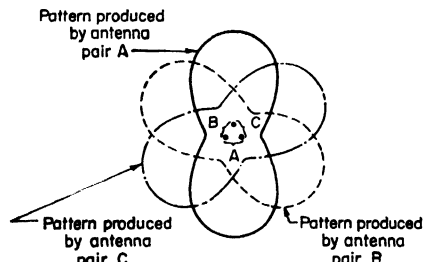



Fig. 12. Navaglobe field pattern. (*IT&T.*)

Navaglobe makes use of three antennas at the corners of an equilateral triangle with a spacing of approximately 0.4λ. Physically, its separation amounts to approximately 4,300 ft at the operating frequency of 100 kc. In its simplest form, the energy from a common transmitter is commutated to one pair of antennas at a time. The resulting space pattern for each pair of antennas resembles a figure of eight with an incomplete null. Three such space patterns result with their relative axis displaced from one another by 120 deg. These patterns are shown in Fig. 12.

Except for the keying necessary to commutate the power to the different antenna pairs, the transmission is essentially c-w. The keying cycle is fundamentally 1 sec long and is repeated continuously under control of an extremely stable crystal oscillator. The cycle is illustrated in Fig. 13. During each 1-sec cycle there are four subintervals. During three of these subintervals, which follow in close succession, the three different combinations of antenna pairs are energized. In the fourth interval, one antenna is excited with an r-f energy differing by about 200 cps from that of the normal frequency which simultaneously energizes a second antenna. This transmission is omnidirectional and serves two purposes. The displaced carrier frequency serves to synchronize a commutator in the receiving equipment which makes possible the identification of the successive signals. In addition, the beat between the two carriers furnishes distance information. A total band width of

20 cps, centered at each of the two carrier frequencies, is all that is required to receive both bearing and distance information.

A second version of the Navaglobe-Navarho ground station makes use of a series of r-f amplifiers located at the base of each of the three towers. An accurate r-f source located in a central location drives the amplifiers. At this central location, keying and control of the amplitude and phase of the components of the signal are accomplished. Relative phase is maintained to 0.05 deg and amplitude to 0.1 db by measurement of the fields at the central location and control of the excitation to the amplifiers. Two amplifiers, operating in parallel, are used at the base of each tower to provide an input power of from 30 to 50 kw. When noise conditions are low, or in the event of failure of one amplifier, one unit may be disconnected.

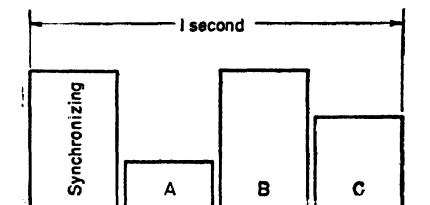
Fig. 13. Navaglobe keying cycles. Pulses *A*, *B*, and *C* correspond to the three patterns shown in Fig. 12. The synchronizing pulse is transmitted omnidirectionally. (*IT&T.*)

The Navaglobe system (in contrast to POPI, which compares phase) determines bearing by comparing the amplitude of the three signals received during each cycle. The relative amplitudes vary with azimuth, as is evident from the space patterns shown in Fig. 12. The receiving equipment utilizes the principle that, if the three magnitudes were added vectorially with a 120-deg space relation, the angle of the resultant would be equal to twice the azimuth angle of the receiver from the transmitter. In practice, this is accomplished by commutating the signals, in a-c form, to the three coils of a sine resolver. The rotor coil, acting as a secondary in the selsyn, has voltages induced in it, the algebraic sum of which is zero when the angle of the coil is set to correspond to twice the azimuth of the receiver from the transmitter. This algebraic sum increases positively or negatively as the coil is shifted about the null point. For automatic direct reading, the resulting voltage may be used as an error signal to control a servomotor that rotates the coil to the null point. For manual operation, a bearing-selector knob attached to the selsyn may be rotated until the indicated-error signal is zero.

It can be noted from the symmetry of Fig. 12 that there are two positions of azimuth, 180 deg opposed, for which the ratio of the three field-strength amplitudes is the same. Hence the cycle repeats twice for a full orbit in azimuth, and the Navaglobe indicator utilizes an expanded scale wherein 360 deg of rotation corresponds to 180 deg of azimuth. The 180-deg resulting ambiguity is relatively easy to resolve in long-distance navigation.

To mitigate against errors produced by noise, the Navaglobe receiver utilizes a wattmeter-type square-law detector and further incorporates a broad-band clipping circuit at receiver input which is then narrowed to an acceptance band of only 20 cps at the output. The circuits also lend themselves readily to postdetector integration. Since each Navaglobe cycle requires 1 sec, a reading may be obtained by noting the position of the wattmeter detector after any number of cycles desired. The equipment is initially designed to provide a reading at the end of every 5 or 15 cycles at the operator's selection. Postdetector integration greatly smoothes and averages the resulting bearing and, under very high noise conditions, integration times of as long as 3 min may be desirable.

A Navaglobe station is capable of furnishing distance-measuring service without modification. In the air-borne equipment, the only modification required is the addition of a second translator unit connected to the output of the receiver. This translator consists of an extremely stable reference oscillator (having a stability of the order of one part in a billion), a count-down circuit, and a phase meter. As mentioned previously, the 200-cps beat, resulting from the simultaneous transmission of two carriers from the ground equipment, provides the basis for distance measurement. The wavelength at 200 cps is approximately 811 nautical miles. Thus, 1 deg of phase difference between the received 200-cps note and that of the local air-borne oscillator is equivalent to a distance of approximately 2.25 nautical miles. To determine distance, therefore, it is only necessary to measure the phase between the arriving note and that generated by the local oscillator. Assuming the phase-measuring system to be accurate to within 1 deg, it can be expected that the system will have an accuracy equivalent to ± 2.25 nautical miles plus any error associated with drift of the air-borne oscillator. The system requires manual adjustment of the phase meter to correspond to the known distance between the aircraft and the ground station at take-off, or over any known check point. At 811 nautical miles from the station, the indi-

cator will have completed a full cycle and will record zero distance. As the aircraft progresses further, it becomes necessary for the pilot to add 811 nautical miles to the reading of the meter. This procedure is followed for any subsequent 811 miles to the maximum range of the equipment. Since these distances are large, there is little possibility of confusion due to positional uncertainty, and therefore no ambiguity resolving circuitry is included.

Summer daytime tests of the Navaglobe system, radiating 6 kw of power, have indicated bearings up to 1,800 nautical miles over both land and sea. These same transmissions, during summer nighttime propagation, have given ranges of 2,000 nautical miles over land and 2,600 nautical miles over sea. Average bearings over a period of about 15 min were found to be within 0.5 deg of the true values. In the presence of severe thunderstorms, bearings within 4 deg of true value were obtained.

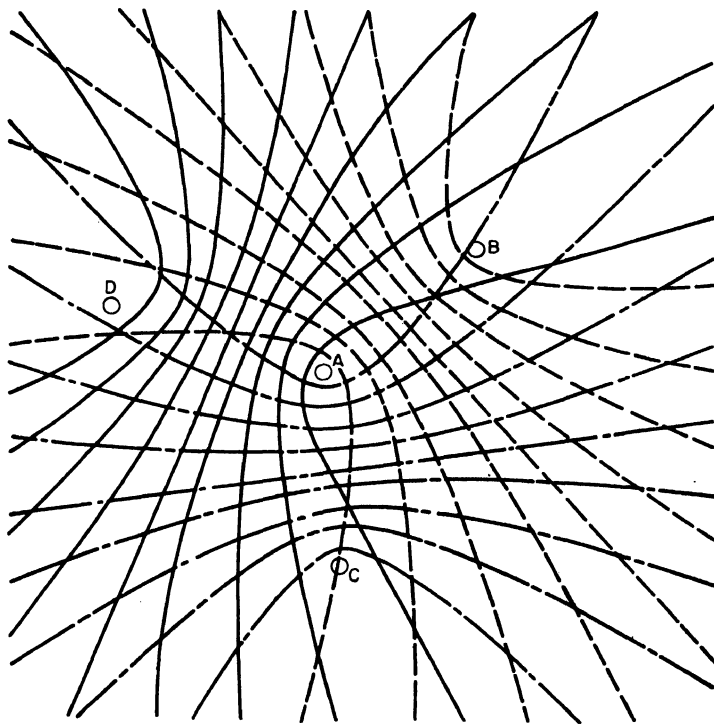



FIG. 14. Decca hyperbolic patterns.

28. Decca. This system, a development of the Decca Navigator Company of London, England, has been proposed for immediate use, but its development continues along the basic principles described here. Decca[22] is a hyperbolic system making use of c-w phase comparisons in much the same manner as POPI. POPI, however, uses a short base line between antennas so that the lines of position will be essentially linear with a maximum phase difference of 180 deg, whereas Decca uses extended base lines (approximately 70 miles) to increase accuracy. Although phase difference between a pair of stations would yield a line of position, Decca is planned as a system that produces a fix; and accordingly it utilizes three stations placed at the corners of an equilateral triangle. The Decca hyperbolic patterns are shown in Fig. 14. A master station is located at the center of the triangle and all measurements are based on phase difference between this central master station and the outer slave stations.

To resolve the problem of measuring the phase difference of two radiations at the same frequency, Decca uses a different frequency for each radiator but each frequency is some integral submultiple of a frequency common to all transmissions. As an example, one of the chains in use employs the following frequencies.

1. Master station is 14.16 × 6 = 85 kc.
2. Slave station *A* is 14.16 × 5 = 70.8 kc.
3. Slave station *B* is 14.16 × 8 = 113.3 kc.
4. Slave station *C* is 14.16 × 9 = 127.5 kc.

The exact values of the multiples are unimportant as long as they are capable of relating master and slave to a common frequency and are integral; for it is on the common frequency that the phase measurements are made.

Another problem that Decca has solved uniquely is that of the measurement of phase difference beyond ±180 deg. Decca employs a cumulative counter which can be set at any point at departure and which will thereafter indicate the total number of complete phase shifts.

The distance between Decca stations makes it necessary that the radiation at each point be generated by an independent transmitter. The radiated frequency is generated by receiving the master station and utilizing its frequency as the source of slave-station frequencies. This procedure assures synchronization between stations. It is further necessary that both master and slave maintain phase to within 1 deg. This is done by receiving the common harmonics of the slave and master stations and applying them to a phase discriminator which controls a reactor tube to correct the phase of the slave transmitter.

The Decca mobile receiver utilizes a phase-measuring system that is independent of the intensity of the voltages being measured. Furthermore, the phase meter has a free movement for a full 360-deg rotation and can therefore be geared to other dials that will add the total number of revolutions and record phase difference to any desired value. A fix is determined by the intersection of two lines of position established by two pairs of transmitters.

The ambiguity problem has been solved by a system of coarse-zone and fine-zone identification. The coarse-zone patterns are displayed on a meter having a sector-shaped pointer that indicates an area of 60 deg. On the face of this same meter are six pointers driven by a common shaft that is operated from the fine-zone identification. In operation, the lanes are identified as red, green, or purple lanes, and in an area that extends to 300 miles, there are 1,058 red lanes, or 1,350 green lanes, or 820 purple lanes. By an extra transmission of the master station, the number of these positions is reduced to about 45 and these are widely separated so that their ambiguity may be resolved by normal navigational means. Selecting the Decca lanes from within the 45 coarse zones is done by a second auxiliary transmission. The ambiguity resolving circuits in the mobile receiving equipment are put into effect by a change in frequency of the master station of 60 cps.

The accuracy figures quoted by Decca are shown in Table 2.

Table 2. Distance and Accuracy of Decca

Distance, miles	Accuracy, yd	
	Day	Night
50	10	20
100	20	50
200	40	200
300	100	500
500	200	
1,000	500	

29. Delrac and Dectra. These systems were recognized by the fifth session of the Communications Division of ICAO and were listed as developments to be encouraged. Both systems are the proposals of the Decca Navigator Company[23,24] and employ techniques closely associated with those which have already been described for the Decca system.

Delrac is a hyperbolic system operating in the 12- to 16-kc band, and may consist of one master and one slave station or as many as three masters each with one slave station, making a total of up to six stations. All stations transmit on the same frequency. Phase lock, however, is required only between a master and its slave. To avoid interference, the transmissions from the various stations are not present at the same time. In the aircraft receiver, a commutating system similar to that described for Navaglobe is used to connect the output of the receiver to crystal oscillators that store the phase of the master frequency. A measurement of the relative phase of a pair of these oscillators then produces the decometer deflections corresponding to the hyperbolic lines of position.

The ambiguity of the Delrac lines of position is resolved by additional transmissions from the master and slave stations. The operation of the equipment may best be understood by reference to Fig. 15. Let it be assumed that F_m equals 12 kc and n equals 3. Then $F_1 = F_m[1 + (1/n)] = 16$ kc and $F_2 = F_m[1 + (1/n^2)] = 13\frac{1}{3}$ kc. In the receiver, F_m from the A station controls the phase and frequency of oscillator A. Similarly, F_m from the B station controls oscillator B. The phases of oscillators A and B are compared in a Decca discriminator and displayed on a decometer. A phase difference of 360 deg corresponds to a distance of about 15.5 miles. Accordingly, although an accurate position indication is obtained, there will be an ambiguity every 15.5 miles. The ambiguity is resolved by extracting a beat note resulting from the mixing of the other frequency and by frequency division. This provides an ultimate frequency of $1\frac{1}{3}$ kc, for which a phase

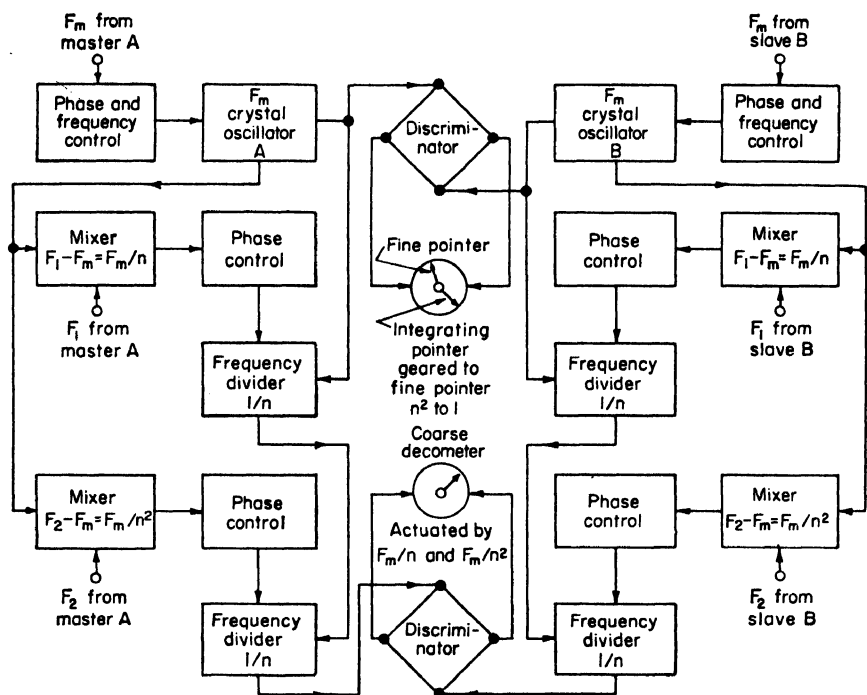



FIG. 15. Delrac receiving system.

difference of 360 deg corresponds to a distance difference of 300 km. There is small likelihood of positional confusion of 300 km. However, if necessary, it is possible to incorporate additional circuits to make use of F_3 and compare frequencies of $1\frac{9}{27}$ kc.

Dectra. In contrast to Decca and Delrac, which are intended to provide essentially omnidirectional service, Dectra is designed to provide service only along a narrow predetermined route. For this purpose, a master and slave station are located about 80 miles apart on a line perpendicular to the route to be traveled. The master and its slave are equally distant from the route and one pair of stations is located at each end of the route. At any one time, track position is determined by the use of only one pair of stations, the second pair being used over the second half of the route.

In operation, the frequency of each master station is the same as its slave and each pair of stations transmits for periods of about 10 sec. The stations at opposite ends of the route, however, differ in frequency. For example, the master and slave at one end may transmit on 85.1 kc while the master and slave at the other end may be on 84.915 kc. Just as a station is about to terminate its transmission, its frequency is changed by 60 cps to cause

the other station of the pair to begin transmitting, and to cause certain switching functions to occur in the receiver.

In the receiving equipment, the phase of the transmitted frequency of the master station is stored by a crystal oscillator and is compared with the phase of the slave station. The equiphase lines of position are hyperbolic but, because the spacing between the master and slave station is small, most of the useful lines have small curvatures and may be regarded as radials originating at the bisector of the line connecting the master and slave stations.

Transmissions from stations at opposite ends of the route are employed to determine distance along the route. For this purpose, frequencies from the stations at opposite ends of the route are connected to a mixer where they beat to produce 185 cps. At the time of take-off, the range indicator is adjusted to read the known distance, and as the transmission from the distant master is received, the range indicator is automatically actuated through action of the distance circuitry.

30. Radux. Radux, a system of long-distance navigation suggested by Dr. J. A. Pierce,[25] employs very long base lines between stations. Radux transmissions are supplied by three or more transmitting stations located far apart in an arrangement similar to that employed by loran. The l-f transmitters are modulated at about 200 cps. The carrier frequencies need not be synchronized, but it is necessary that the modulation frequencies be carefully synchronized. The transmitters are keyed sequentially for a period of about 0.6 sec with a gap of 0.2 sec between transmissions. The sequence is stopped for about 0.8 sec before being repeated. The gap in transmission is used to identify the three transmissions.

Since the keying cycle requires in excess of 3 sec, the Radux receiver employs a very narrow keying-cycle filter that is expected to reject most of the noise. The output of the filter drives a commutator that connects the output of the receiver successively to three storing circuits. In one embodiment of the receiver, these storing circuits may be accurate 200-cycle tuning forks. A phase detector and servo are used to correct the phase and frequency of the forks so that they are in synchronism with the incoming signals. The outputs of three or more sets of tuning forks are then continuously compared in phase in circuits of the type that have already been described for the l-f omnidirectional range or Decca.

31. High-frequency Direction Finding from Ground Stations. Many types of ground-station direction finding have been devised for use in conjunction with air-borne transmitters operating in the l-f, m-f, h-f, v-h-f, or u-h-f bands. Two or more such DF installations, located at widely separated sites, will permit a fix on the location of an aircraft. Position determination can also be accomplished by means of a single DF operating in conjunction with range-finding equipment.

The useful range of an h-f or l-f DF depends upon many variable factors, including the frequency of transmission, the location and power of the transmitter, the time of day, and the season of the year. Although these factors do not apply to the same extent in v-h-f and u-h-f transmissions, these higher frequencies are limited to line-of-sight range. The accuracy of a DF is dependent upon such factors as atmospheric conditions, polarization, scattering, diversity, and site effects.

The h-f DF was not considered for standardization as an international long-distance aid by ICAO at its meeting in the fall of 1946, but it was recommended for use in several regions, notably the South Atlantic and Middle East. It is highly improbable that the device will receive recognition as an international standard long-distance aid; however, it serves a very useful purpose as an emergency ground aid to long-range navigation. In rescue work, it will probably continue to play an important part for many years to come.

32. Direction-finding Antennas. Antenna systems used with DF equipment are necessarily directional in their characteristics but may vary radically in their design. The simplest system makes use of a vertical-loop antenna supplemented by a nondirectional antenna, the combination producing a cardioid pattern. This system eliminates the 180-deg ambiguity which is present when a loop alone is used. Such an antenna can be made small enough so that it has found wide application on special-purpose portable equipment. Loop antennas are capable of good performance on

ground-wave signals but are adversely affected by reflected sky-wave signals, the latter inducing voltages in the horizontal sides of the loop and thus introducing bearing errors or instability. These polarization errors are greatly minimized by employing the Adcock antenna or the spaced-loop antenna.

33. The Adcock Antenna. In 1919, an English patent[26] was granted to Adcock for a DF antenna system having, as its main purpose, the elimination of polarization error. Fundamentally, the principle involved the use of elements sensitive only to the vertically polarized component of the radio wave. Various forms of this antenna system are shown in Fig. 16, where the antenna shown at *A* is the original or U-type form. The other forms shown are modifications by later inventors for the purpose of further reducing the polarization error. The basic theory of the Adcock array depends upon the figure-of-eight pattern which is formed when the spacing between two elements is small compared with a wavelength. If the spacing is greater than $\lambda/2$, a multilobed pattern is produced which will result in ambiguities.

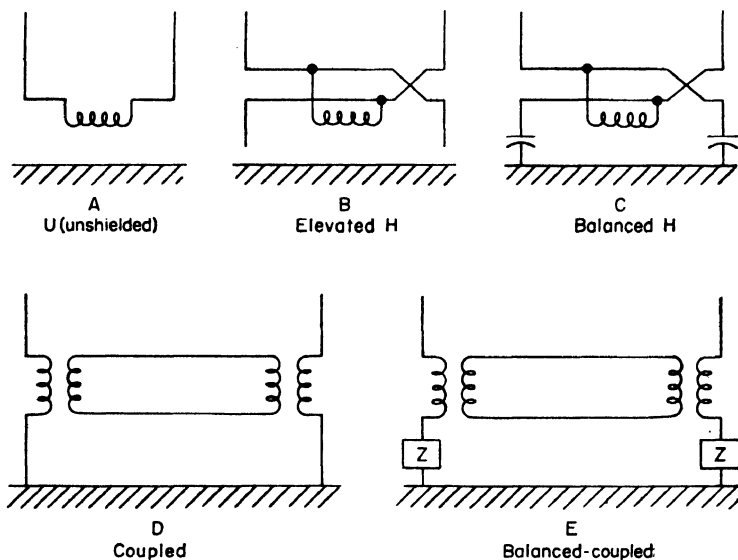



Fig. 16. Common forms of Adcock antenna systems.

Practical Adcock arrays utilize two crossed Adcocks associated with an inductive-type mechanical goniometer, or its electrical counterpart, and form the equivalent of a rotatable single Adcock. This arrangement is shown in Fig. 17. The mechanical goniometer utilizes a rotary search coil which is coupled to the two stator windings crossed at right angles, with each fed by one Adcock. The resultant flux in the goniometer will be the vector sum of the flux produced by each field coil. Accordingly, if the search coil is rotated until the voltage induced is maximum, it will be in a position (with respect to the plane of the goniometer coils) corresponding to that of the plane of propagation of the radio waves with respect to the Adcock antennas. For greater sensitivity, the bearing is obtained by rotating the search coil until a null is received, and noting the position of the coil at this time.

The electronic goniometer may take a variety of forms. In one form, two balanced modulators supply a two-phase modulating voltage to the Adcock system. The two Adcock antennas feed a common receiver, but the signal voltages corresponding to the two antennas may be individually recovered in the receiver output by reference to the two phases of the special modulating or scanning signal. In a second form, the north-south antenna is fed to the receiver through a balanced modulator operating at one audio frequency, while the east-west antenna is fed through a second balanced modulator operating at a different audio frequency. The resulting two side bands are amplified equally, passed

through filters, and are used to represent the amplitude and phase of the voltages induced in the two Adcocks by the incoming wave. With either form, sense of the bearing may be introduced by proper combination with the output of a nondirectional antenna.

34. The Spaced-loop Antenna. Another means for minimizing the effects of abnormally polarized waves is the spaced-loop system devised independently by R. A. Weagant in the United States and C. S. Franklin in England, and later investigated and developed by T. L. Eckersley.[27] The principle of operation is illustrated in Fig. 18. The loop antennas are connected in phase opposition and the pair is then rotated about a vertical axis through point P to seek a null. Analysis of the system indicates that, regardless of the value of the incident and polarization angles, the spaced-loop output voltage goes through a null when the directions of propagation are contained in the planes of the loop antennas. This differs from the Adcock antenna in that the

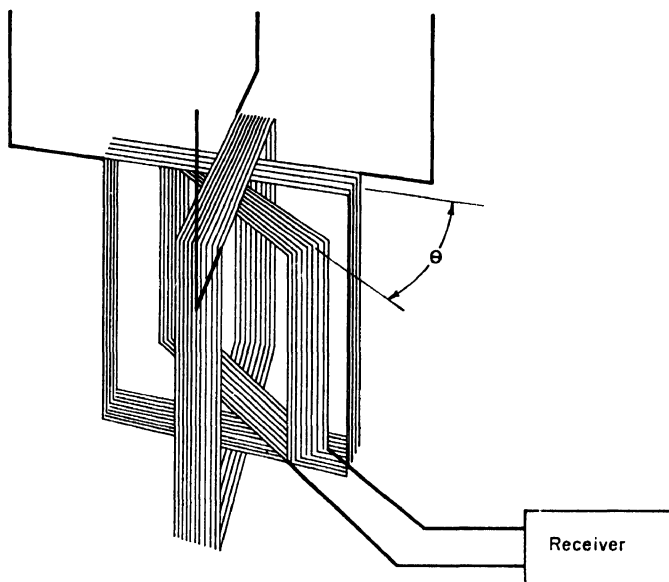



FIG. 17. Goniometer for Adcock system. Outer coils go to antennas and bear a 90-deg relation to each other. Inner coil rotates.

spaced-loop DF makes no attempt to reject any component of the impinging wave. Measurements made by Eckersley concluded that the polarization error was only 0.3 deg. When the direction of propagation is normal to the plane of the loops, additional nulls are produced which are referred to as "loop nulls." Under these conditions, it is often necessary to provide a means for resolving the ambiguity.

35. Direction-finder Systems. All DF stations require a directional receiving antenna, a radio receiver, and a bearing indicator. To determine fixes, the control center is also required to coordinate the operation of a station network. Complexity of equipment may vary from lightweight portable units to large fixed installations, depending upon the nature of service and upon the requirements for elimination of polarization errors. The effectiveness of any DF system may be materially increased through proper instrumentation which gives simultaneous visual indication, through panoramic techniques, of all frequencies received. This assures that any desired signal will not be overlooked if slightly off frequency.

Ground-station DF can employ either aural or visual indications, or both. Aural indication of the bearing is obtained by rotating the antenna or goniometer until a

null point is reached, and has the advantage that bearings may often be taken under high noise conditions where visual means often fail. Visual indications may be presented in the form of a minimum reading on a meter or, more generally, as deflections on the screen of a cathode-ray tube graduated in degrees around its outer edge.

With cathode-ray displays, the goniometer must be rotated at high speed, mechanically or electronically, to obtain a stable pattern on the screen. Modern DF favors the use of electronic scanning with units of the antenna system remaining stationary. Cathode-ray indications may take a large variety of forms but are usually arranged to give a lobe or line of indication on the tube corresponding to the direction of arrival of the received wave. Typical indications are shown in Fig. 19. One such indicator

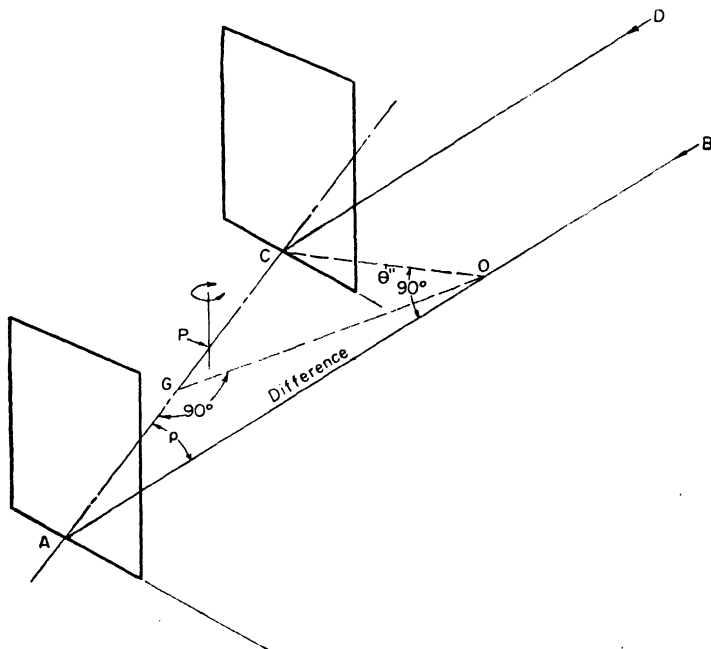



Fig. 18. *BD*, direction of incoming wave. *CO*, perpendicular from *O* to *C*. *AO*, difference in distance traveled by radio wave in reaching the two loops.

is used by the Signal Corps SCR-291. Several variations of this device were produced, with different designations, and over 1,000 units were manufactured. On this equipment, the bearing appeared as an elongated figure-of-eight pattern (a twin-leaf pattern) and a line through the points of the leaf indicated the bearing. When the sense antenna was momentarily connected, the leaf pattern changed to a cardioid to resolve the ambiguity.

Fixed errors may occur in any crossed-Adcock system, because of the geometry of the antenna spacing, and are referred to as *octantal errors*. Mechanical goniometers also introduce an octantal error, but these fixed errors may be removed by calibration or by a compensating distortion in the form of presentation. Within a statistical probability of 50 per cent, average bearing errors with DF equipments are about 1.5 deg on ground waves and 3.5 deg on sky waves, respectively. A spaced-loop DF developed by United Airlines prior to the Second World War produced, for many readings, less than 4.5 deg of error for 95 per cent of the bearings and less than 1 deg error for 50 per cent of the bearings. Bearing data are normally classified as class A, class B, or class C for accuracies within ±2 deg,

±5 deg, or more than ±5 deg, respectively; and fixes are similarly classified for accuracies within 5, 20, and 50 miles or more, respectively. Methods of statistical plotting are in use to improve bearing accuracy by averaging or weighing the data produced by the direction finder.[28-30]

36. Loran. The loran system was developed[31] during the Second World War by the Radiation Laboratory in the United States and derives its name from the words "long-range navigation." Its primary purpose was to meet the requirements of military operations and to offer some protection against jamming and enemy use. The choice of frequency, compromising between best accuracy and the longest range, fell between 1,700 and 2,000 kc.

The basic operating principle[32] of loran is the measurement of time difference in propagation between two widely separated transmitting points and the receiving equipment. Since the velocity of propagation of radio waves is a constant, this is

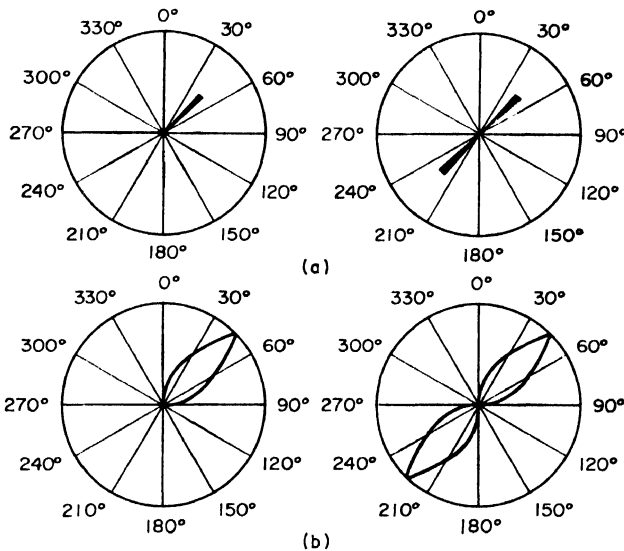



FIG. 19. Typical cathode-ray bearing indications for ground-station direction finders.

equivalent to a measurement of the difference in distance. Loran transmissions are in the form of pulses initiated by widely separated sources, and the difference in arrival time at the receiver is an indication of the distance difference. The locus of points of constant distance difference is a family of hyperbolas; therefore loran is a hyperbolic system.

Athough loran falls short of meeting the requirements of ICAO, its high accuracy and range were recognized by the Special Communications Technical Committee of that organization who, in their meeting during the fall of 1946, urged "that standard Loran, which is now providing important coverage over certain vital ocean areas, be retained and extended or supplemented as may be necessary to meet traffic requirements." By 1954, more than 25 chains employing 53 stations were in operation, furnishing service over a great area of the world, including principally the North Atlantic area and the Central and Southwest Pacific Ocean area.

37. Loran Principles. The loran principle of operation is illustrated in Fig. 20. A chain of stations, providing two sets of hyperbolic position lines, is necessary for a fix. The chain may consist of four stations of which two are masters and two are slaves; or one master may control two slaves, or one slave may be pulsed by two masters. The

maximum practical base line between master and slave is 600 miles, since reliable ground-wave reception (for synchronization of slave with master) is not obtained at greater distances, even over all water paths. A 300-mile base line is conventional. Figure 21 shows a typical arrangement of three stations, with A as the master, for establishing intersecting loran position lines over an area.

Maximum accuracy is obtained when the spacing of lines is closest together and when intersecting lines are perpendicular. Since the longer the base line the greater

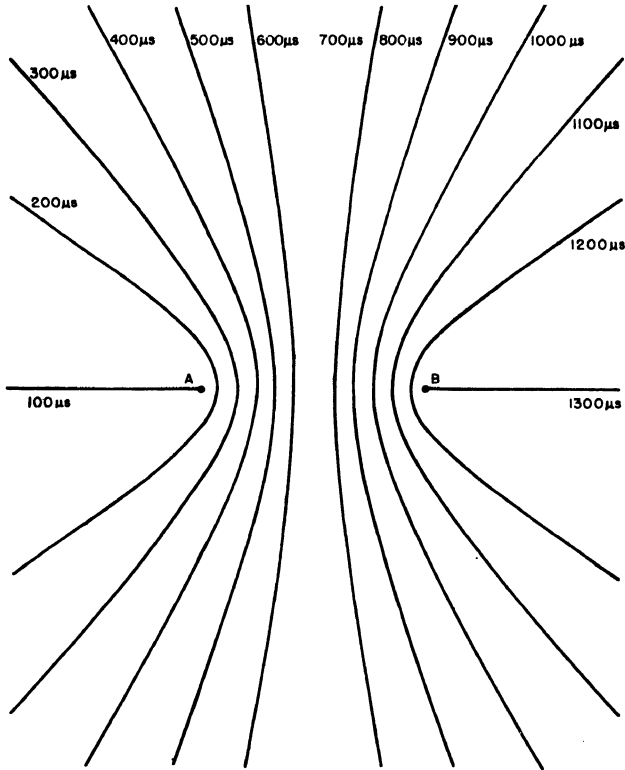



FIG. 20. Family of hyperbolic lines of position formed by two antennas 600 µsec apart with the A pulse delayed 700 µsec behind the B pulse.

will be the area for which the loran hyperbolas will be approximately straight and parallel, two properly disposed loran pairs can be used for high-accuracy coverage of an extensive area.

Considering a pair of loran stations, the master station transmits pulses of about 80 µsec duration and of about 70 to 100 kw peak power. Loran designs utilize a repetition rate of 25 or 33⅓ pulses per sec. Pulses from the master station are received at the aircraft after an interval representing the time of propagation. Pulses from the master station are also received at the slave station after some other interval characteristic of the base line used. These received pulses cause the slave station to transmit other pulses at the same repetition rate but with a fixed delayed time. The slave transmission is thus synchronous with the master transmission. Pulses from the slave station arrive at the aircraft after a time interval proportional to the distance from the slave to the aircraft. The aircraft therefore receives two series of pulses and the time interval between these, as measured by suitable time-marker pips, represents the desired time or distance difference.

Loran time intervals are represented in a system of symbols as follows:

T = indicated time difference
L = recurrence interval
γ = coding delay
C = velocity of radio propagation
T' = true time difference
D = absolute delay
β = time taken for pulse to travel from master to slave

All times are measured in microseconds.

Assume that a pulse is transmitted from the master station at $t = 0$. The pulse will arrive at the slave station at a later time $t = \beta$ and the slave pulse will be radiated from the slave station at a still later time $t = \beta + L/2 + \gamma$ which is equal to the absolute delay D. At any receiving point, the true time difference between the master and slave pulses will be $T' = \beta + L/2 + \gamma \pm X/C$ (where X/C is the difference in actual travel time of a radio wave from the master and slave stations to the point). However, owing to the fact that the time difference $L/2$ is automatically taken care of by the presentation of the two pulses in the receiving display, the indicated time difference as used by the navigator will be $T = \beta + \gamma \pm X/C$.

For standard loran, β is usually 1,611 μsec (300 miles) and γ is normally fixed at 1,000 μsec.

38. Loran Receiving and Indicating Equipment. While the specific techniques employed in various receiver-indicator models differ, fundamentally each consists of a conventional radio receiver, a stable oscillator that serves as the source of highly accurate time information, and a cathode-ray oscilloscope on which the received signals are displayed.

FIG. 21. Typical arrangement of three loran stations to form two loran facilities or pairs.

A nonautomatic modern receiver-indicator weighs approximately 40 lb and consumes approximately 190 watts. To receive loran fixes, the receiver must be capable of operating on four different r-f channels. These channels are 1,750, 1,850, 1,900, and 1,950 kc. The superheterodyne receivers utilize crystal-controlled oscillators. The appropriate operating channel is chosen by manipulation of a channel switch. The receiver consists of one stage of r-f amplification, a converter, three i-f stages operating at 1,050 kc, a detector, and one video stage. The total band width, at 6 db down, is 80 kc. The sensitivity is such that a 10-μv signal produces full deflection of the oscilloscope. The image-frequency rejection of the receiver, as well as its response to all spurious signals, is 60 db below the response to the desired signals.

An unusual feature of the receiver is a differential gain control which operates in synchronism with the incoming signal recurrence frequency, and thereby permits different amplification of each of the two signals received from different loran pairs. This feature permits presentation of equal amplitudes on the cathode-ray indicator even with incoming signals having a ratio as high as 500:1.

Selection of the station pairs operating on the same frequencies is done by changing the rate of the oscilloscope sweep. Although all signals will be present on the oscilloscope face, only those signals which are synchronous with the chosen sweep rate will remain stationary. The basic loran repetition rate is varied so as to permit eight choices; therefore the loran

system has provision for receiving and differentiating between 64 different pairs of operating channels at any one time. The first basic rate is 25 pps and is varied by $\frac{1}{16}$ pps per step until the rate of $25\frac{1}{2}$ pps is reached on the eighth step. Similarly, the high basic rate of $33\frac{1}{3}$ pps is varied in steps of $\frac{1}{6}$ pps per step until $34\frac{1}{6}$ pps is realized.

The indicator unit contains the necessary circuits to perform all the timing functions of the equipment with the required precision. The basic timing medium is a precision crystal-controlled 100-kc master oscillator with an extremely high short-term stability. Manual means are provided to vary the frequency by approximately 200 ppm, which permits recycling the oscillator until the timing of the receiver-indicator is in exact step with the pulses received from the transmitting stations. Through the medium of its timing circuits, the indicator provides a sequence of precise timing markers, spaced at convenient intervals, to facilitate measurement of time sequences with a basic accuracy of the order of ± 1 μsec.

The exact character of these timing marks differ with different receiving systems, but the method of measurement involves the sequential rotation of a manual function switch through various positions (from 5 to 7). A typical loran display is shown in Fig. 22 and the operation of the function switch, in one model of a receiver, would be as follows.

FIG. 22. Presentation and interpretation of loran signals. (*U.S. Coast Guard.*)

In position 1, a slow-sweep generator is used wherein each trace lasts for half the recurrence rate; that is, if the ground-station recurrence rate was 25 then the sweep would last for 20,000 μsec and then be repeated. The divider chain also provides a square-wave signal for the vertical plates of the oscilloscope, thereby deflecting the horizontal trace so that one appears above the other. Two vertical deflections are then individually observed on each of the traces representing the received signals from a loran pair of transmitters. A fine frequency adjustment is then made to position the vertical deflections properly and to stop all movement. With the function switch in position 2, fast sweeps are initiated (1,400 μsec) providing a generous expansion of the signals on the face of the oscilloscope. This step also permits differential control of the receiver gain so as to set the two magnitudes to be equal and provides for more accurate time superposition. In position 3, the separation between traces is removed and by use of a fine delay the signals are accurately superpositioned. Thus, the amount of the delay between the starting of the two traces now represents the time between the signals. The remainder of the measurement does not refer to the signals, but only to the locally generated timing signals. Accordingly, in position 4, the radio receiver is no longer in use and the slow-sweep circuit is again reinsti-

tuted in a manner similar to function 1, except that timing marks are superimposed on the display. The presentation appears as in Fig. 23 and the operator counts the number of whole 1,000-μsec marks between the start of the trace and the variable-marker signal. In the example shown, the reading is approximately 6,000 μsec. Finally, position 5 permits a fine measurement of the delay. The fast sweep is again in use but, to permit greater accuracy, the trace is nonlinear. During the first 100 μsec the cathode beam moves rapidly, but during the remaining 1,300 μsec the voltage time slope of the sweep is less steep and the beam moves across more slowly. This last measurement permits the final determination of time interval to within ±1 μsec.

In addition to the five positions described, the receiver normally incorporates a sixth position used for homing. In this operation, the delay controls are first preset to values corresponding to the line of position that is desired; then the aircraft is flown so that the signals on the fast sweep are superimposed.

To obtain a fix, two pairs of stations must be selected and a function switch must be rotated through its positions twice while making the various adjustments. In latest models of loran receiving equipment, further advanced circuits have been incorporated to increase the ease of reading and decrease the time required to obtain a fix. These equipments convert the pulses to sinusoidal waves having fundamental frequencies equal to the recurrence rates of the pulses, after which phase shifters are used to obtain the equivalent superposition. The phase shifters are geared to a number display which indicates the delay directly. The latest model permits making readings on two stations simultaneously.

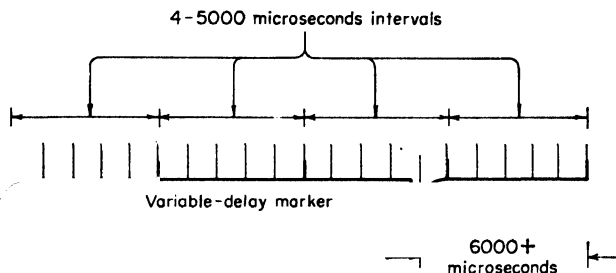



FIG. 23. Appearance of cathode-ray display with function switch in position 4. The reading is 6,000 + μsec.

Loran charts and tables are published by the United States Hydrographic Office for use aboard ships and aircraft. Each chart shows lines of position for several pairs of transmitting stations, the family of lines for each pair being printed in a distinctive color. These lines are spaced at regular intervals, usually 20 μsec, and every line bears an identification symbol denoting the r-f channel, the basic pulse rate, and the specific pulse rate. The loran tables contain, in tabular form, essentially the same information as the charts.

39. Loran Range and Accuracy. Standard loran signals propagate via ground wave and sky wave, and both types of propagation produce usable signals. Table 3 indicates the signal field strength required to give usable readings for various seasons, latitudes, and times. The table indicates a series of computations for propagation over sea water and shows a maximum range that varies from 400 to 850 miles depending upon conditions of propagation. For transmissions over ground, the maximum distance may be one-half to one-third of the values shown, depending on the conductivity of the soil.

The table values are for propagation via the ground wave only, but loran pulses also travel via the ionosphere and, at night, usable signal strengths are obtained to a maximum of about 1,500 nautical miles. This range is essentially the same over land because only the ground in the immediate vicinity of the transmitter is significant in sky-wave propagation.

Various errors are inherent in the elements of the loran system, such as lack of perfect synchronization between the transmitted pulses (normally ±0.5 μsec), errors in sky-wave contamination, and errors in charts and tables. Furthermore, the area of

Table 3. Range of Loran System

Latitude	Season	Day		Night	
		Required signal, μv/m	Range, nautical miles	Required signal, μv/m	Range, nautical miles
Equatorial	....	25	550	250	400
Middle	Summer	5	700	50	500
	Winter	1	850	10	650
Arctic	Summer	1	850	10	650
	Winter*	1	850	1	850

* The range in winter in quiet regions is limited by the finite sensitivity of the loran receivers.

best accuracy is where the hyperbolic lines are closest together. The positional accuracy may therefore vary over a wide range. In general, aircraft receivers permit readings to about 1 μsec, which, when added to the station synchronization error of about 0.5 μsec, may produce an error of 1.5 μsec. This corresponds to an error of ±788 ft from the proper line of position.

40. SS Loran. This form of loran was an attempt to produce a highly accurate, very-long-range system operating on standard frequencies. Four stations were employed with their base lines intersecting at nearly 90 deg, and all readings were made within the area bounded by a line connecting the four stations. The base lines were 1,000 to 1,400 nautical miles in length. Hence, it was possible to synchronize the chains only via the sky wave, from which the initials SS (sky wave synchronized) were derived. For this reason the system operated only during the night hours. Over most of the coverage, the intersection between position lines from two pairs of stations was over 70 deg; therefore, accuracy was good despite the loss due to uncertainty in the propagation factor. Measured errors varied from 0.6 to 5 nautical miles depending on distance and angle between directions to the stations. No special equipment was required for receiving SS loran in the aircraft. The ground equipment, however, was modified by use of timers incorporating extended delays.

41. Low-frequency Loran. Low-frequency loran was conceived during the last year of the Second World War through the joint effort of the Signal Corps and the Radiation Laboratory of MIT. The system operated at a frequency of 180 kc and comprised four stations located at the corners of a 1,000-mile square to provide a useful coverage of 1,790,000 square miles for day and 1,370,000 square miles for night operation with an rms distance error not exceeding 5 miles. A pulse length of 300 μsec was used, and because of this, it was not possible to match pulses to read the time difference with the same accuracy as with standard loran. The lower accuracy of time-difference readings was compensated, to a certain extent, by the use of longer base lines and by exhibiting the r-f wave on the oscilloscope rather than its rectified envelope. With this system, it was possible to match the r-f cycles and thus increase accuracy.

The summer daytime range with low-frequency loran was 1,000 nautical miles over land and 1,200 nautical miles over sea. The summer nighttime range was more than 1,500 nautical miles. A large number of readings indicated median daytime accuracies of between 1 and 4 nautical miles with distances to 1,500 nautical miles. Similarly, nighttime accuracies of between 1 and 7.5 nautical miles were indicated with distances to 1,500 nautical miles. In both cases, accuracies decreased sharply with distance beyond 1,500 nautical miles.

More recently, receivers have been designed to permit cycle matching of the r.f. and thereby improve the accuracy. This system has been termed "loran B," and receivers of this type are becoming available. The term "loran A" has been applied to the standard pulse-matching loran.

Because the low-frequency loran receiver required a broad band to pass a pulse, there was a question regarding its successful operation in tropical regions. In 1946, ICAO recommended that additional tests be conducted; but before it was possible to do so the International Telecommunication Union at its 1947 conference held in Atlantic City established a frequency of 95 ± 5 kc for use by low-frequency loran and similar systems. In view of this recommendation, it was necessary to redesign all the equipment before tests could be made. By 1954, test results were at hand and at the fifth session of the Communications Division of the ICAO it was recommended that no further consideration be given to the use of this system.[33] However, system development was continued, predominantly for tactical purposes, and this has led in recent years to a type known as "loran C." Loran C is a development of the Sperry Gyroscope Co. and constitutes the only l-f hyperbolic navigation system which is free from sky-wave contamination. This is accomplished by utilizing only the first few r-f cycles of the pulse which always arrive at the receiver prior to the sky wave. A cycle-matching technique is then used to provide an improved order of accuracy. Although still experimental in nature, loran C has shown promise of providing accurate navigation data at ranges in excess of 1,500 miles because of its ability to discriminate against the sky wave.

42. Electronic Pilotage Equipment. Electronic pilotage equipment is the name applied to apparatus heretofore referred to as air-borne radar. The term is intended to apply to air-borne radio equipment that presents in the aircraft a presentation of the terrain below, so that the aircraft may be directed from point to point by an observation of the landmarks appearing on the radio map.[34] Radar of this type was first developed as a means for permitting a night fighter to close in on an enemy bomber and later for target identification in bombing missions. It was first applied in surface detection in about 1941 to locate submarines. Following the war, the equipment was used solely for navigational purposes.

Early equipments were lacking in detail, but were used generally in the following manner. Knowing the starting point, the navigator plotted a course to be flown and noted the time when the first prominent radio-map landmark would appear, and its relationship to the course flown by the aircraft. Data taken from this observation were used to compute a course correction. This dead-reckoning procedure was followed between successive landmarks. With this method, electronic pilotage equipment was useful in spite of its lack of clear detail. Definition was later improved by the use of 3-cm wavelengths instead of the 10-cm then in use.

Equipment was also designed that could be used with ground *transponder beacons*, that is, stations that receive the transmissions from the air-borne equipment and reply by means of a ground transmitter. When used in this manner, the range is limited to line of sight. This application is accordingly not feasible in the long-range en-route zone and the following discussion is limited to unassisted radar.

43. Principles of Electronic Pilotage Equipment. In visual pilotage, the area under observation is illuminated by the sun, and the reflected radiations impinge upon the human eye to form an image which is interpreted in terms of terrain and landmarks. With radio equipment, it is not practical to illuminate the entire area under observation since this would require the generation of prohibitive amounts of power. Instead, the radio energy is concentrated in a beam 3 to 6 deg in total width, and the area is sequentially illuminated by scanning or rotating the beam. At any instant, therefore, reflections are received from a limited area only and, if a 3-deg beam is used, the maximum azimuth definition possible is of the same order. The distance to the terrain objects is determined by transmitting the energy in pulses and measuring the time required for the signal to travel to the object and return. These pulses are of the order of 1 μsec; hence the maximum axial resolution is of the order of 983 ft.

Equipment has been constructed utilizing wavelengths as short as 1 cm, but because of atmospheric absorption, a wavelength of 3 cm has been found to be more practical.

In early radar equipment, one antenna was used for transmitting and a separate antenna used for receiving the reflected energy; however, in all modern equipments the same

antenna serves both purposes and eliminates possible errors arising from the incorrect positioning of two separate antennas. The principle of electronic pilotage equipment is shown in Fig. 24. The main components of the system consist of a pulsed transmitter, a highly directive antenna system, a receiver, and display apparatus. In Fig. 24, a pulse generator of 1 μsec duration excites a modulator which, in turn, impresses the pulse on the transmitter. Hence the transmitter, operating at 10,000 Mc, generates short bursts of r-f energy concentrated in a narrow beam by the antenna. The direction of the antenna is continuously varied at a rate of about 180 deg per sec. The same pulse used to trigger the transmitter is connected to the deflecting coils of an oscilloscope, after being converted to a saw tooth by the sweep generator. Accordingly, the cathode-ray beam is caused to deflect from the center of the tube to the outermost edge in exactly the time that is required for the radio wave from the antenna to travel to the maximum desired distance and return. The line thus formed can be directly calibrated in distance. As the antenna changes azimuth, the oscilloscope coils are rotated by a mechanism that assures coincidence between

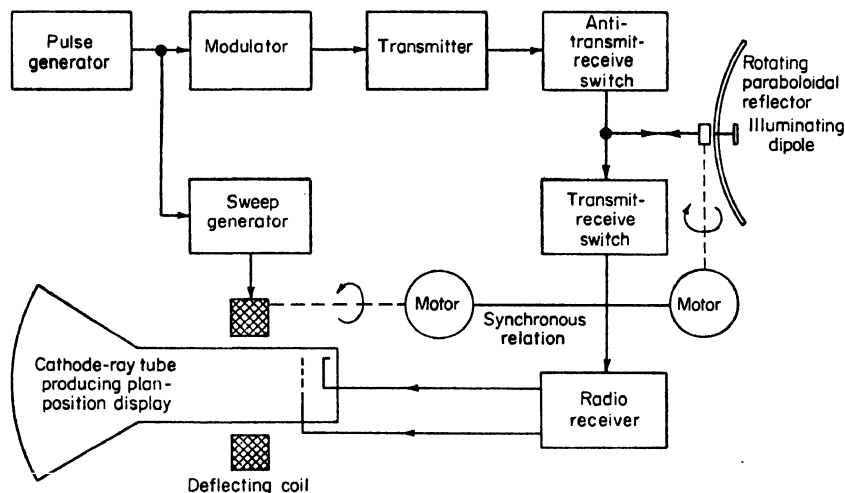



FIG. 24. Block diagram of electronic pilotage equipment.

the position of the deflection coils and the direction of the antenna. A map-type presentation results from which the azimuth of any landmark from a given reference, and its distance, may be directly observed.

In choosing the pulse-repetition rate, the maximum desired range must be considered. A second pulse cannot be transmitted until the first has had time to return by reflection from an object located at the greatest desired range. For a 100-mile range, the maximum p.r.f. would be 920 per sec.

To facilitate the observation of distance, calibration marks equating tube distance to true distance are produced on the face of the tube by electronic means. The definition is limited by the angular beam width, since the minimum width of any spot, regardless of how small the reflecting obstacle, must always be the beam width of the antenna. In distance, or along the axis of the beam, the definition is limited by the pulse width.

Although it may be concluded that electronic means are poorer in definition than visual means, this is compensated to some extent by a much more accurate knowledge of distance and size than is normally offered by visual means. Of greater importance is that electronic pilotage equipment can provide information at night or when weather conditions do not permit visual observation, and can do so at much greater distances than is possible by eye alone.

44. Characteristics and Performance. One of the most common electronic pilotage equipments constructed during the Second World War had an installed weight of about 190 lb, with a power consumption of approximately 1 kw. Subsequent designs have

reduced weight and power consumption, and one of these for commercial-aircraft use weighs approximately 60 lb with a power requirement of approximately 600 watts.

Electronic pilotage equipment in service is capable of detecting cities at a distance of 25 to 50 miles at altitudes of 10,000 ft. It is also capable of detecting recognizable land features such as bays at about the same distance. Mountains may be detected at approximately twice this range depending on their elevation. Typical accuracies of recognizable landmarks are approximately 2.5 deg in azimuth and within 10 per cent of the distance. These accuracies, however, are not the limiting factors in the usefulness of electronic pilotage equipment. For example, over the sea where there are no readily observable landmarks, the equipment is without value except for the recognition of small islands. A typical terrain presentation is shown in Fig. 25. This

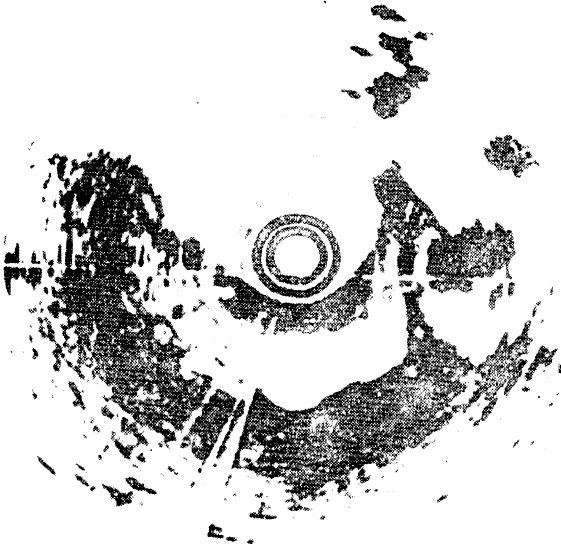



FIG. 25. Plan-position display located in aircraft flying in the vicinity of Sandusky, Ohio. (*Courtesy of CAA.*)

presentation may be compared with the map of the same area shown in Fig. 26. Even in cases of terrain with easily recognizable features, it is necessary to interpret the radar map carefully to ensure that features are not being identified incorrectly.[35]

The chief advantage of electronic pilotage equipment is that it is self-contained and therefore has no range limitation. Accordingly it is useful in areas where no other navigational facilities are available. As with any pilotage equipment, its use requires accurate maps.

45. Electronic Aids to Dead Reckoning. Dead reckoning, a term derived from "deduced calculation," is defined[36] as the procedure of determining present position by the addition of one or more vectors representing known courses and distances from some previous fix. This process can be extended to project the known position toward the desired destination. Normally, the process is applied in the following manner. Knowing his present location the navigator determines from a map the distance and direction to the desired destination. The direction to the destination is termed a course and is held by reference to a directional reference such as a magnetic compass. Knowing the speed of the vehicle, the navigator then determines the length of time required to cover the distance to the destination. After the calculated time has elapsed, the navigator may reasonably assume that he is at or near his destination.

This process, when used in an aircraft, has one important exception. Since the pilot is interested in his position with respect to the surface of the earth, the movement of the air mass (winds) with respect to the earth must be taken into consideration in making the calculation. In this case, it cannot be assumed that the heading and course are the same. By taking into account a vector representing the direction and magnitude of the wind, the pilot may compute a "drift-correction angle" that can be added to or subtracted from the "course angle" to obtain a heading to be steered so that the desired track is made good. Although data regarding the magnitude and direction of the wind are usually available, this information is seldom reliable. Winds may change rapidly in magnitude and direction with respect to time, location, and altitude. The pilot therefore follows the practice of making a normal dead-reckoning calculation, but

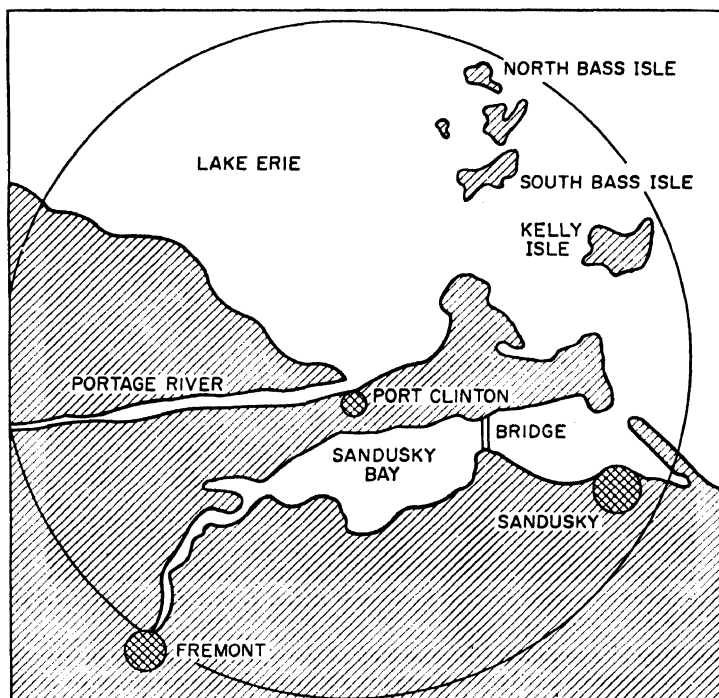



FIG. 26. Map of area represented by PPI display of Fig. 25.

then corrects it from time to time by observation of his position. This position is determined by observation of recognizable landmarks or by obtaining radio fixes. In this sense, therefore, all radio aids may be considered as electronic aids to dead reckoning. This section is limited to those devices which supply parameters usable in making the dead-reckoning calculation (but do not necessarily give lines of position or fixes) and which are self-contained in the aircraft.

46. High-altitude Radio Altimeter. This altimeter should not be confused with the low-altitude altimeter sometimes referred to as a *terrain-clearance indicator*. Developed during the Second World War to increase the accuracy of high-altitude bombing, the high-altitude radio altimeter saw far more use as a navigational aid than in the application for which it was originally designed. Since the end of the war, this device has become extremely popular, particularly in transoceanic flights.

The high-altitude altimeter is a simple radar which determines distance in one direc-

tion only, namely, directly below the aircraft. Its principle of operation is similar to that described for the electronic pilotage equipment.

In dead-reckoning calculations, the device is used in the following manner. It is known that the direction and magnitude of winds are a function of pressure areas. In the Northern Hemisphere, the winds blow clockwise around a high-pressure area and counterclockwise around a low-pressure area, the converse being true in the Southern Hemisphere. The indication of the barometric altimeter is a function of both altitude and sea-level air pressure; therefore, by determining the absolute altitude through the use of the radio altimeter, the sea-level barometric pressure can be determined. The change in barometric pressure over a period of time describes the pressure system prevailing along that portion of the flight path. While flying at constant-pressure altitude, the radio altimeter is used to determine true altitude at convenient intervals of time. The velocity of the wind component perpendicular to the heading of the aircraft may then be computed by the following formula: $W = K(D_2 - D_1)/A$. In this formula W is the wind velocity in knots, D_1 is the difference in feet between the radio and barometric altimeter noted at the first reading, D_2 is the same difference at the second reading, and A is the air distance expressed in nautical miles. The factor A is derived by multiplying the true air speed by the time between readings. K is a constant varying with latitude and given in navigational tables intended for this purpose.

The principal components of a radio altimeter are a pulse-type transmitter, an antenna system, a receiver, and an indicating device. High-altitude radio altimeters utilize a frequency of 440 Mc. This frequency is sufficiently high to permit generation of short pulse widths and yet allow the use of relatively standard components and techniques.

To avoid the necessity for transmit-receive switches, separate antennas are used for these functions. These antennas consist of simple dipoles mounted $\lambda/4$ below the undersurface of the aircraft fuselage or wings. The $\lambda/4$ spacing reinforces the radiation downward with a gain of approximately 3 db. To minimize drag, the dipoles are normally mounted with their long dimensions parallel to the line of flight. Since the mutual coupling is high for this configuration, a portion of the fuselage must be located between the antennas to act as a shield and reduce the direct transmission between them. If, however, the antennas are located "end on" the coupling between them is essentially zero and no shielding is necessary.

FIG. 27. Appearance of the altitude indication on the cathode-ray tube.

For altimetry, a peak power of from 5 to 10 watts is all that is required; hence the transmitter can utilize receiving-type vacuum tubes. The r-f energy is normally pulsed by two sources, a crystal-controlled source generating a frequency of 98.356 kc and an LC oscillator source generating a frequency of 9.835 kc. The higher frequency produces greater accuracy but presents ambiguities every 5,000 ft which are resolved by the low-frequency source. The ratio of frequencies indicates that no ambiguities are present for the low-frequency transmission within 50,000 ft, which is the service range of the equipment. Pulse widths of the order of 0.25 μsec are used, which correspond to a distance accuracy of approximately 250 ft. This, then, is the theoretical minimum distance at which the altimeter could be used. It should be noted that the pulse-repetition rates are high compared with other types of radar equipment. This is permissible since the maximum distance to be measured is 50,000 ft, which is short compared with expected performance in other types of radar.

A typical altimeter display is shown in Fig. 27. The presentation takes the form of a concentric circle on which two lobes protrude outward toward the edge of the tube.

The first lobe indicates the initial transmission and is located directly upward, while the other lobe, representing the reflected pulse, is at a position which is a function of altitude. A transparent scale calibrates the position with respect to altitude.

The circuit is arranged so that the higher-frequency sweep will always be a little smaller in diameter and thus be readily identified by the operator. A control marked "times one" or "times ten" serves to select the proper sweep. With the fast sweep, a p.r.f. of 98.356 kc is utilized whose period corresponds to the time required for a pulse to travel 10,000 ft. Accordingly, the sweep is calibrated to a maximum of 5,000 ft. For this altitude, the indicator will have made a complete rotation and be coincident with the direct pulse. Because the sweep is continuous, there will be no distinction between altitudes differing by 5,000 ft. To resolve this, the slow sweep is used which utilizes a p.r.f. of 9.8356 kc, which is then read against a 50,000-ft calibration. By reading the scale at the leading edge of the echo during the fast sweep, it is possible to read to within 50 ft regardless of altitude, so long as the aircraft is higher than approximately 500 ft. Considering all errors, the accuracy of the altimeter can be considered to be 50 ft ± 0.25 per cent of the altitude.

A widely used model of a high-altitude radio altimeter weighs 35 to 45 lb installed, and requires 135 watts of power.

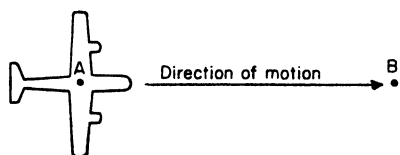
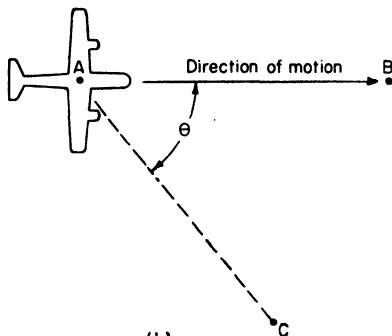
47. Doppler Systems. The Doppler drift is a well-known phenomenon in sound and light whose fundamental application, using radio waves, is shown in Fig. 28. In this figure, a radio transmitter is assumed to be located on a moving vehicle A traveling at a speed S in the direction indicated. At some observing point B, the received frequency F_r will differ from the transmitted frequency F_t because of the motion of the vehicle in the following manner.

$$F_r = F_t\left(1 + \frac{2S}{c} + \frac{S^2}{c^2}\right)$$

Normally, the last term is so small that it may be neglected. If the receiving point is located to one side of the direct line of motion, such as point C, then the relative velocity of the vehicle with respect to the observing point will be a function of the angle θ. Under this condition, the following formula holds.

FIG. 28. (a) Transmitter at A travels toward B. The total number of waves passing B is the sum of those for the vehicle at rest plus those carried by because of the vehicle motion. (b) Velocity of vehicle C is a function of the cosine of θ.

$$F_r = F_t\left(1 + \frac{2S\cos\theta}{c}\right)$$

If the observing points reflect energy back to the transmitter, the Doppler frequency is derived by beating a small amount of the original transmitter energy with the reflected energy.

As an example, if a wavelength of 3 cm is employed and the vehicle is traveling at a rate of 300 mph, then the value for $\theta = 0$ gives a Doppler frequency of 8,940 cps. If, however, the transmission from the vehicle is directly at right angles to its line of motion ($\theta = 90°$), the Doppler frequency will be zero. In essence, the Doppler frequency is directly proportional to the rate of closure between the transmitter and the reflecting point.

In the application of Doppler techniques to a self-contained navigation system, two air-borne antennas can be imagined mounted at right angles. The first of these directs a beam of energy along the axis of the aircraft, but pointing slightly downward. The other antenna produces a beam at right angles to the first. By detection of Doppler frequencies along these two axes, both the forward and sideways velocity of the aircraft can be determined. Hence, the actual direction of motion of the aircraft with respect to its heading becomes known.

A single antenna may also be utilized if the antenna is rotated and the components read at different times. This permits a simple modification of existing electronic pilotage equipment. A servosystem is utilized to rotate the antenna with a positional accuracy within 0.25 deg. An additional equipment, weighing not more than 15 lb, then furnishes means for providing drift determination by the Doppler method.

Over reasonably level terrain, drift readings may be obtained with probable errors of the order of ⅓ deg.

As a self-contained navigation system, Doppler techniques may be used to obtain ground speed directly. The method was originally suggested by J. J. Hibbert and N. W. MacLean.[37] Such a device, in conjunction with an appropriate computer, can constitute a self-contained automatic dead reckoner which has the capability of determining present position with respect to a starting point.

48. Inertial Navigation. Inertial navigation was primarily conceived and developed for military applications such as the guidance of long-range missiles and long-range bombing missions. In such applications, its significant advantages include the impossibility of enemy detection (since no radiations are emitted), its independence of ground facilities, and its invulnerability to enemy jamming. Although these advantages are much less important in commercial operations, it is likely that inertial navigation will find future use in civil operations, particularly for long-range overseas flights. Many significant advances have been made in the art of inertial navigation in the past several years, but most of the details of this work, as of this writing, are covered by military security regulations.

An inertial navigator can be considered to be a form of dead-reckoning device, since position is determined with reference to the starting point by proper processing of the data that represent past movements of the vehicle. Basically, the inertial system determines displacement by measurement of the accelerations of the vehicle relative to the earth. In its simplest form, imagine a vehicle equipped with an accelerometer capable of measuring the acceleration of the vehicle along the line of motion. Integration of the acceleration value will yield the velocity of the vehicle. A second integration, operating on the velocity value, will then yield distance traveled. Hence, if the direction is known, the position of the vehicle with respect to the starting point will be established.

If the vehicle is to have unlimited freedom of motion, it is obvious that the acceleration components must be measured in more than one direction. A minimum of two and, in most applications, three accelerometers are required for this purpose which measure the mutually perpendicular components of acceleration. Since it is vital that the accelerometers measure only the acceleration forces on the vehicle and ignore gravity, it becomes necessary to mount the accelerometers on a gyro-stabilized platform so that the direction of gravitational forces will remain fixed and known. By the double integration of the mutually perpendicular acceleration components, the distance traveled by the vehicle in any direction (including the vertical) can be computed and thereby establish an absolute position if the starting point is known.

The principal components of a typical air-borne inertial navigator consist of the accelerometers, a gyro-stabilized platform, a spherical trigonometric computer, and a device to generate special corrections. These special corrections include principally the fact that the earth is not a perfect sphere and the Coriolis effect generated by the rotation of the earth.

At the initiation of the flight, it is necessary to erect the stabilized platform so as to be horizontal and align the accelerometer assembly so that the sensitive axis of the

north-south accelerometer is pointing in the north-south direction. If absolute position with respect to the earth is desired, the latitude and longitude of the starting point and destination must be inserted into the computer. As the flight progresses, the double integration of the accelerometer signals provides the distance traveled in the north-south and east-west directions, respectively. These distances, when added to the starting coordinates, indicate the present latitude and longitude positions.

The system just described makes no allowance for the fact that the earth is spherical and that gyros maintain their position with respect to space rather than with respect to gravity. To provide this additional information and avoid the spurious measurement of gravity acceleration, a device similar to a pendulum is required. An ordinary pendulum, however, is of little use since it would be adversely affected by normal accelerations of the aircraft, particularly for high-speed planes and missiles. This problem is overcome by using the principle first described by Dr. Maximillian Schuler of Germany in 1923. This principle, alternatively called the *Schuler pendulum* or the 84-min pendulum, makes it possible to establish the direction of gravity in a moving vehicle, regardless of its accelerations. In his original description, Dr. Schuler pointed out that a violent deflection of a pendulum in a moving vehicle can be avoided only if the length of the pendulum is equal to the radius of the earth. For example, if the bob of the pendulum were at the center of the earth, it is obvious that the point of suspension can be moved at will without changing the position of the bob. Because the earth is not a perfect sphere, the length of such a pendulum must be varied slightly with latitude. However, this is a second-order effect which is easily compensated for in modern inertial systems.

It is obviously impossible to use a pendulum of the form described above, but the same effect can be obtained by any physical pendulum having a period of oscillation equal to a string pendulum of the required length. This period of oscillation turns out to be approximately 84 min. The effect of an 84-min pendulum is synthesized in the aircraft by a feedback system which utilizes both the gyro and accelerometer signals.

Hybrid Systems. The greatest challenge to the designer of modern inertial equipment is the design of components of greater precision and with reductions in weight. At the present state of the art, an inertial system may weigh anywhere between 75 and 2,000 lb, depending upon the accuracy desired, the length of time involved, and the intended purpose of the mission. Much work is being done on increasing the sensitivity and accuracy of accelerometers and gyros to make possible more precision and less weight. At the present, these components have proved difficult both in development and in manufacture, and are extremely costly. A pure inertial system of sufficient accuracy to be reliable as a long-range navigation aid still weighs many hundreds of pounds. To overcome this problem, many observers have suggested the use of a "hybrid" system in which the inertial navigator is combined with other forms of navigation. These hybrid combinations are practical with any self-contained or dead-reckoning equipment and are not exclusively limited to inertial navigators. The hybrid equipments can utilize a supplemental radar, celestial observation, or radio ground-reference techniques to correct occasionally the cumulative errors which build up in pure inertial or dead-reckoning systems.

One example of a hybrid system is the use of a Doppler radar to measure the ground speed of the aircraft. This is then compared with the first integration of the inertial equipment, and if any steady-state errors exist a correction is made in the inertial equipment. Such a system can be used even under military conditions since the Doppler radar would be required only for a few seconds at intervals of perhaps 10 or 15 min. Another possibility is the combination of celestial and inertial navigation in which a fix is obtained by the use of a sextant. One such sextant has been developed which provides for automatic tracking of preselected stars or planets by night or the sun by day. This system is useful at high altitudes where celestial observations may always be made irrespective of weather. A third combination is the use of any ground radio-navigation system capable of providing a fix of the desired accuracy. This would avoid the necessity for continuous operation of civil-navigation aids in the event

of a major war, and thereby avoid the use of these aids by enemy bombers. When combined with an inertial-navigation system, aids such as VORTAC or Tacan could be turned on briefly at irregular intervals to enable a pilot to reset his inertial navigator and avoid cumulative error. A similar arrangement could be used with long-range navigation aids such as loran and Navarho. Engineers believe that an inertial navigator of limited accuracy weighing of the order of 30 to 40 lb could be constructed for use in a hybrid system.

In summary, inertial navigators are capable of supplying the following information without radiation of signal and without reference to ground aids.

1. The geographic position of an aircraft in latitude and longitude, or as a function of the distance and direction from any starting point
2. The ground speed and track
3. The distance traveled from the starting point and the distance remaining to the destination
4. The direction of the destination
5. The attitude of the aircraft

As an obvious extension, it should be noted that the gyro-stabilized platform can be used for other purposes such as the control of an automatic pilot, the stabilization of radar antennas, and as a precise reference for heading instruments. Thus, a coordinated instrumentation in the aircraft could save appreciable weight. However, in spite of much progress to date, many problems still exist in the development of accurate and reliable components, and it is generally agreed that inertial navigators for common civil use are many years in the future.

PART 2

THE EQUIPMENT OF THE EN-ROUTE SHORT-DISTANCE ZONE

49. Requirements. The en-route short-distance zone has been defined[38] as that portion of the airspace in which the long-distance en-route zone terminates, and which is served primarily by radio aids spaced at intervals of less than 200 miles. The en-route short-distance zone in turn terminates in, but it does not include, the approach or airport zone. Unlike the long-distance zone, the primary problem of the short-distance zone is that of traffic control. This occurs because of the convergence of traffic in the short-distance zone which rises to a maximum density in the approach zone.

Normal traffic-control procedures in the en-route zone consist in reserving a block of airspace for each aircraft. This procedure may be formal, where clearances are filed with a ground governmental agency, or may be informal where only flight advisory services are employed. Whichever procedure is used, a successful traffic-control system must be capable of making reservations with sufficient rapidity so that the movement of an aircraft into the desired space can be made without delay.

A primary requirement of navigational aids in a short-distance zone is that they shall be capable of providing guidance along the shortest possible route. Accordingly, full airspace coverage is preferred as opposed to discrete courses or airways. This provides maximum air safety in traffic control since the location of the aircraft is known anywhere in the short-distance zone.

A second important requirement, relating to traffic control, is that sufficient accuracy[39] must exist to define adequately the boundaries of the reserved airspace. From another point of view, greater accuracy permits a reduction in the dimensions of the reserved airspace and thereby permits a higher traffic density with safety.

A third important requirement of short-distance navigational aids is that they must furnish guidance at all times, even under adverse atmospheric conditions. In high

traffic densities, loss of guidance for even short periods greatly increases the dangers of collision. To complete the requirements for safe traffic control, rapid and reliable communications is also vital.

In summary, a short-distance navigational aid must provide complete airspace coverage with sufficient accuracy to control safely the expected traffic density, and incorporate, if possible, reliable communication service.

50. The V-h-f Phase-comparison Omnidirectional Radio Range (VOR). Historically, the work which led to the modern VOR was begun in about 1928 by Englund,[40] Evans,[41] and Greig[42] of the Bell Telephone Laboratories. Although their work stopped in 1930, interest in the device was revived by the work of Dr. D. G. C. Luck[43,44] which began in about 1937. In 1940, CAA undertook the continued development of VOR, which extended through the Second World War. The outcome of the development[45] was presented to the RTCA at a meeting in January of 1945.

In 1946 the United States government offered VOR for worldwide use in the en-route short-distance zone at a meeting of the Special Communications Division of what was then the PICAO. The device was accepted by this organization, as indicated by the following excerpt from Annex 10 of the Standards and Recommended Practices of the ICAO: "In localities where conditions of traffic density and low visibility necessitate a short distance radio aid to navigation for the efficient exercise of air traffic control, the standard aid shall be the VOR (VHF Omni-range) of the CW phase-comparison type conforming to the standards contained in paragraph 3.4 of Chapter 3 of this Annex."

51. VOR Principles. The VOR principles of operation are similar in many respects to the CAA l-f omnidirectional range discussed in Sec. 25. In brief, the operation of the VOR may be likened to that of an airway light beacon revolving at a uniform rate. If the beacon makes 1 rpm, and if a fixed red light on the beacon tower is caused to flash as the beam points north, an observer with a stop watch can determine his bearing merely by starting the watch when the red light flashes and stopping it when the beam flashes past him. The azimuth angle measured from north is then directly proportional to the time recorded by the stop watch.

The omnirange functions through the radiation of two separate r-f field patterns generated by a common transmitter. The first, or variable-phase pattern, corresponds to the rotating beam of the beacon, while the second provides the signal reference and corresponds to the flashing red light. These field patterns are produced by an antenna array consisting of four horizontal-loop-antenna elements of the Alford[46] type located at the corners of a square, and a fifth element located at the center. The center element is excited with v-h-f energy in the 112- to 118-Mc range. The r.f. is amplitude-modulated approximately 30 per cent with a 9,960-cps signal. The 9,960-cps signal is, in turn, frequency-modulated with a 30-cps reference signal with an excursion of ± 480 cps. In addition, voice modulation not exceeding 40 per cent may be applied to this same carrier. Accordingly, the central element produces a nondirectional reference field pattern which also serves as a voice channel.

The diagonally disposed pair of elements at the corners of a square are connected to a common feed point by different lengths of transmission lines so that one element is 180 deg out of phase with the opposite element. The radiated pattern is thus a figure of eight. The same arrangement is employed with the other pair of antennas diagonally disposed on the other two corners of the square. The transmitter supplies energy to these two pairs of antenna elements through a special capacitor-type goniometer, the rotor of which is driven at 1,800 rpm (30 rps) in phase synchronism with the reference generator. The goniometer functions like two mechanical side-band generators, suppressing the carrier and feeding two 30-cps modulated side bands (displaced by 90 deg in phase of the modulation envelope) to the two pairs of antenna elements.

Since the modulation envelopes of the individual patterns are in phase quadrature, the resultant produces the equivalent of a rotating field. This field, when combined with the carrier of the central antenna, forms a limacon pattern. At any observing point, the rotating limacon pattern, when demodulated, produces a 30-cps signal which constitutes the variable-phase term. The demodulation of the f-m 9,960-cps tone

produces the reference 30-cps component. With proper calibration of the ground and air-borne equipment, phase comparison of these two components yields the bearing of the aircraft from north.

52. VOR Equipment. The antenna array developed by CAA is illustrated in Fig. 29. The four coplanar loops are mounted at the corners of a square such that the diagonal or spacing between opposing pairs is 112.2 electrical degrees. The four antennas occupy positions approximately 48 in. above a counterpoise while the central antenna is mounted on a pedestal $63\frac{3}{8}$ in. above the counterpoise. With this spacing of 112.2 deg, octantal variation of about 2.5 deg in phase will occur. This error may double if the two side-band voltages differ in amplitude by as much as 10 per cent.

FIG. 29. (a) CAA antenna array using five Alford loops. (b) Plan view of CAA VOR antenna array.

A block diagram of the VOR system is shown in Fig. 30. The essential components are a v-h-f transmitter, a voice and reference modulator, a goniometer, the f-m reference generator, and the antenna system previously described. The c-w transmitter is crystal-controlled and operates in the band from 112 to 118 Mc. Transmitters have been designed with various power outputs ranging from 50 to 300 watts. An important requirement is that the transmitter shall be capable of 100 per cent modulation from 300 to 12,000 cps and that there shall be no phase shift with modulation from 9,480 to 10,440 cps. When voice is used for station identification, it is recorded on film that is moved continuously by a motor, and repeats the identification at intervals of between 10 and 20 sec. A second method of identification is a 1,020-cps tone keyed with International Morse code signals.

The principle of the rotating capacitor-type goniometer is shown in Fig. 31. Two rotor plates, insulated from each other, form half cylinders with the ends tapered. When rolled out flat the shape of each plate is equivalent to double half-sine waves. Two sets of stator plates, each stator consisting of two insulated half-cylindrical shells, are disposed side by

FIG. 30. Schematic of transmitter for VOR.

side about the rotor, so mounted that they are displaced by 90 deg. V-h-f voltage from the transmitter is applied across the two rotor plates and each stator feeds one pair of differentially connected antenna elements. The capacitor goniometer produces two v-h-f outputs proportional to the sine and cosine of the displaced angle of the rotor, respectively.

The reference generator consists of a tone wheel driven from the same motor that rotates the goniometer, thereby ensuring perfect synchronism between the variable phase and reference tone. The teeth of the tone wheel are made of magnetic material and are swept by the pole piece of a permanent magnet. The teeth have variable spacing to generate in the coil a nominal frequency of 9,960 cps modulated at a 30-cps rate. This subcarrier, already frequency-modulated, is then applied as amplitude modulation to the carrier. An alternate method is to generate a synchronous 30-cps tone which then frequency-modulates an independent 9,960-cps oscillator. The tone-wheel method, however, offers the advantage of increased stability and reliability.

An alternate antenna system of novel design was developed by the Federal Telecommunications Laboratories[47] for use with VOR. The purpose of the development was (1) to minimize the difficulty of installation and adjustment by reducing the number of transmission lines to the antenna system, (2) to reduce octantal error by physically (rather than electrically) rotating the field pattern, and (3) to reduce polarization error by a structure of special design.

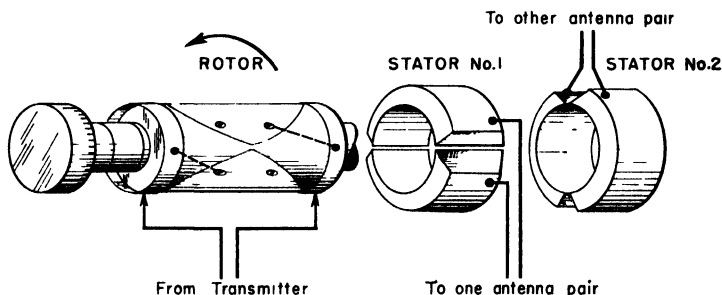



FIG. 31. Rotating-capacitor type of goniometer for v-h-f omnirange.

Briefly, the structure consisted of two cylindrical cagelike structures made of vertical rods. The lower structure, approximately 6 ft high, is the source of both the reference-phase and variable-phase field patterns. The upper structure, approximately 12 ft high, serves the purpose of suppressing the unwanted or vertical polarization component and has no electrical connections to the transmitter. The diameter of the system is approximately 30 in. Electrical connections within the structure are made by means of a circular wave guide. A dipole of $\lambda/10$, internal to the lower structure, is rotated by a small synchronous motor to obtain the desired pattern rotation.

53. The VOR Receiver. Omnidirectional-range receivers designed for commercial service are crystal-controlled and provide preset channels every 100 kc. Since some receivers provide as many as 280 channels, these receivers incorporate crystal-saving schemes to eliminate the necessity for employing a crystal for each channel. As typical examples, the Federal Telephone and Radio type 145 receiver employs a triple superheterodyne to achieve the desired performance with 21 crystals for 140 channels. The Collins 51-R receiver is a double superheterodyne and employs 34 crystals to produce 280 channels.

The sensitivities of commercial receivers are such that an input signal of 3 μv, modulated 30 per cent with a 1,000-cps tone, produces 200 mw of output with a signal-to-noise ratio of at least 6 db. AGC maintains the signal constant within 3 db for signals that vary from 5 to 50,000 μv.

The primary function of the receiver is to recover the reference and variable-phase 30-cps components so that they may be compared to yield the bearing of the receiving point with respect to the omnirange station. A further function is to provide course indications suitable for flying a course along the radial corresponding to that bearing. A

typical circuit to provide this service is shown in Fig. 32. The reference signal is separated from the complex audio signal by a 10,000-cps filter. After limiting, frequency demodulation, and further filtering, the reference 30-cps tone is fed to a phase splitter that produces two outputs of 90 deg phase relation. These outputs are then fed to the stationary coils of phase shifters. One of these phase shifters is driven by an induction motor to provide an automatic indication, while the other is operated manually and is called a *bearing selector.*

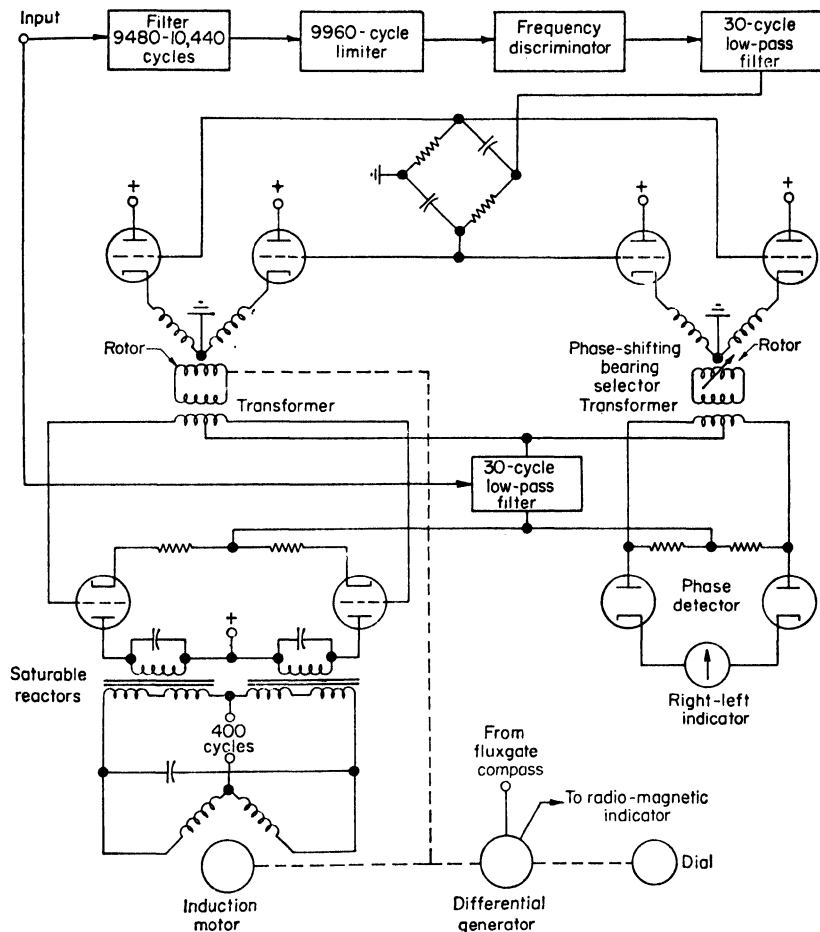



FIG. 32. VOR bearing-indication circuit.

The 30-cps variable-phase component is recovered by amplitude demodulation and, after filtering, is fed to a phase detector whose reference is obtained from the manually operated phase shifter. The output of the phase detector operates the vertical needle of a cross-pointer instrument. When the bearing of the aircraft from the ground station is equal to the bearing that has been set on the bearing selector, there is no deflection of the needle. By following right or left deflections of the needle, a radial course may be flown to or away from the VOR station. In general, the sensitivity is set so that a variation in bearing of 30 deg causes a full-scale deflection from right to left.

The output of the motor-driven phase shifter goes to a second phase detector whose output controls two saturable reactors connected in series with the windings of a split-

phase induction motor. The motor is therefore energized until the reference and variable-phase signals are in phase. The position of the shaft is read on a dial and represents the relative bearing between the VOR station and the receiving equipment.

To provide an RMI (*radio magnetic indicator*) (Sec. 73), a differential generator is utilized in the circuit. The rotor of the generator is connected to the output of a flux-gate compass. The card on the RMI indicator, calibrated from 0 to 360 deg, rotates as a function of the magnetic bearing of the aircraft. This bearing may be read against a fixed lubber line located at the top of the instrument. The needle is rotated in accordance with the output of the differential generator. The compass card and needle rotate together for a change in aircraft heading, assuming that there has been no change in the bearing of the aircraft from the range station. For a change in bearing, but not in heading, the needle alone moves. Therefore, if the pilot desires to fly directly to the VOR station via the shortest route, he alters the heading of the aircraft until the range needle points to the lubber line.

Without an RMI indicator which exhibits bearing from 0 to 360 deg, a possibility of ambiguity exists. Using only the manual phase shifter and the cross-pointer indicator, a zero indication is obtained at two points representing reciprocal courses. To resolve this ambiguity, a to-from indicator is normally supplied to distinguish between the direct and reciprocal courses. When flying over the station, this indicator deflects from "to" to "from."

In addition to providing the indications required for short-range navigation, modern VOR receivers are also capable of receiving transmissions from runway ILS localizers of either the equisignal or phase-comparison type. Moreover, the same cross-pointer indicator also serves during the approach to indicate the localizer and glide-slope paths.

A number of receiving antennas of different designs[48] are in use, the simplest being a horizontal V-type dipole with an 80-deg apex angle and fed by a coaxial line. Another common type is the U or *ram's-horn* antenna which consists of two members with streamlined cross sections forming a horizontal U. These are mounted on a pedestal somewhat less than $\lambda/4$ above the fuselage. The pedestal includes a matching transformer. Other designs[49] are available for high-speed aircraft which feature very low aerodynamic drag. A preferred mounting position for the VOR antenna is atop the vertical stabilizer, where attitude effect and propeller modulation have been found to be minimum.

54. Range and Accuracy of VOR. The range of VOR is limited by radio line-of-sight considerations and, even with aircraft of high-altitude capability, does not extend beyond 200 miles. For aircraft of the commercial type, the range can be considered to be of the order of 100 miles.

The bearing accuracy of the VOR system is limited by six main sources of errors, which are listed below.

1. Ground-station octantal error
2. Receiver-indication error
3. Polarization error
4. Vertical-pattern effects
5. Site effects
6. Terrain effects

The first three errors can be classified as equipment errors while the last three can be considered propagation effects.

The ground-station maximum octantal error is 2.5 deg for the standard CAA antenna system. Although the Federal antenna should theoretically produce no octantal error, in practice it exhibits an error of 1.75 deg.

Receiver-indication errors are produced from such factors as unequal phase shifts in the receiver channels for the variable-phase and reference-phase signals, errors in the phase detector, and inaccuracies in the resolver. These errors exhibit a slow systematic variation with bearing and may vary with individual receivers and with aging of components. In typical receivers, the maximum value of these errors may be between 0.5 and 1.0 deg.

Polarization errors occur when the vertical field component of the ground transmission is excessive. These errors have been termed *course push* or *attitude effect*. These names stem from the fact that bearing changes are noted if the aircraft is banked, so that its receiving antenna becomes more efficient in receiving vertically polarized radiations. The error results since the vertically polarized component, being of an uncontrolled nature, is not likely to contain the same modulation information as the desired horizontal component. To reduce polarization error to the order of 0.1 deg the vertical field reception should be suppressed approximately 60 db below that of the horizontal field. Typical flights performed by the CAA have shown attitude errors ranging between 1.5 and 2.6 deg.

The field strength produced by any antenna system mounted above ground is dependent upon the sum of two components, the direct component, and the component derived from reflection from the ground. At some points in space, these components are out of phase and produce a signal minimum. The number of such lobes is a function of the electrical height of the radiator above the ground plane. When flying in the vicinity of these minima, momentary loss of signal results. When the signal coming directly from the range station reaches a low value, other signals reaching the receiver via reflection from nearby areas (and which are normally insignificant) assume a major importance. The result is a possible departure of the bearing indication from its true value, and these errors are classified as *site* or *terrain errors*.

Site effects and terrain effects may vary widely in error depending upon the site chosen. The VOR system produces indications within accuracies of ±1.5 deg on radial courses free from serious reflections. In general, properly installed VOR facilities can be expected to produce total accuracies of better than ±5 deg.

55. The VOR Monitor. The CAA does not permit the use of equipments as navigational aids in blind flying unless the equipment is adequately monitored and unless a fail-safe alarm indication is presented to the pilot. On the ground, the monitoring system consists of a remote receiver and phase-comparison unit. This receiver, normally similar to the air-borne design, is housed in a weatherproof case and located approximately 200 ft from the station. The monitor measures all significant quantities in the transmission, and a deviation beyond acceptable limits of these quantities energizes an alarm and causes stand-by equipment to be actuated. In the aircraft, absence of the complete signal (or any component of the signal) actuates a flag alarm.

56. Distance-measuring Equipment. An adaptation of wartime radar, DME enables a pilot to determine his distance from a transmitting station. The device saw its beginning as intercept equipment on aircraft early in the Second World War. During the use of this early equipment, it was found that a beacon (or transponder) located on the ground could be used in conjunction with air-borne radar to furnish valuable navigational information. These transponders became necessary since the v.h.f. then employed for intercept purposes did not permit distance reading at low altitude by pure radar reflection. One of several early systems was known as **BABS**, and was derived from "blind-approach beacon system."

The usefulness of distance indications led to the development of special systems which used the combination of air-borne radar interrogators and ground transponders exclusively for navigational purposes. Experience gained illustrated the advantages of a distance-direction presentation (polar navigation) in which the frame of reference was a point on the earth. It was evident that such a system permitted the convenient generation of an unlimited number of fixed flight paths. From this background, engineers of the Wright Air Development Center of the United States Air Force therefore urged[50] the development of a distance-measuring equipment that could be used in conjunction with the v-h-f omnidirectional radio range being developed by the CAA.

The advantages of distance-measuring equipment were recognized by technical representatives of other nations, and a recommendation[51] for such equipment was made by the Special Radio Technical Division of the ICAO that met in Montreal in November of 1946.

In 1949, ICAO recognized u-h-f DME in its Standards and Recommended Practices

with the following paragraph appearing under the recommendation for short-distance radio aids. "It is intended that UHF-DME . . . will become a basic component of the VOR . . . at the earliest date practicable, to be added to all VOR which have been installed before that date."

During the interval, successful distance-measuring equipment had been constructed operating in the 200-Mc range. Although the ICAO standardized on this frequency band, it left room for the use of 1,000-Mc equipment by including a developmental specification for DME operating in the 960- to 1,215-Mc range.

CAA policy is to add DME to all VOR and ILS stations. This equipment, operating in the 960- to 1,215-Mc band, uses both frequency and pulse coding to identify the individual channels. This system, normally referred to as the "civil DME" or "CAA-DME," is described below. Another existing system is the Tacan-DME (sometimes called "DMET"), whose significant difference is that channeling is

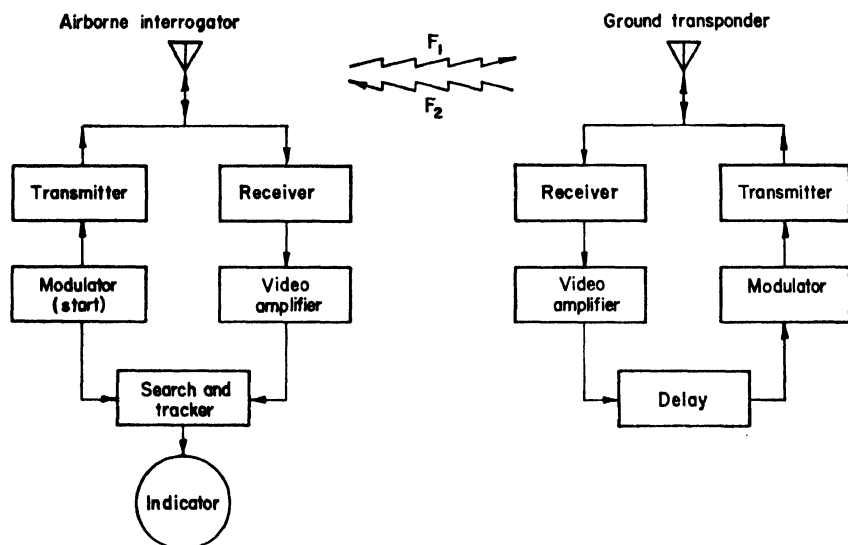



FIG. 33. Distance-measuring system consisting of air-borne interrogator and ground transponder.

accomplished by frequency separation only. For details of Tacan, see Sec. 71. Since the two systems utilize the same frequency range but different channeling, they were incompatible for use in a common ground environment. Accordingly, in the fall of 1956, the Air Coordinating Committee recommended that the Tacan-type channeling be standardized for use in the United States. When combined with the VOR, this has been termed VORTAC. The description that follows, except for channeling and certain specific equipments, applies equally to both systems since the principles of operation are identical.

57. Principle of DME. The DME principle is illustrated in Fig. 33. The system consists of air-borne equipment known as an interrogator and ground equipment referred to as a *transponder* or beacon. The process begins in the air-borne interrogator by the generation of omnidirectionally radiated pulsed transmissions at a frequency F_1. The pulse-repetition rate is relatively low and seldom exceeds 150 per sec. These pulses are usually generated in pairs for reasons that will be explained later. The pulses are received by the ground beacon and, after a fixed delay time, are used to trigger the transmitter of the ground transponder. These response pulses are also radiated omnidirectionally but on a frequency F_2. The response pulse is then received by the air-borne equipment and the time of reception is measured relative to the time of initial transmission of the interrogator pulse. This time

interval is directly proportional to the distance between the aircraft and the ground beacon. Hence, the basic measurement is the round-trip radio-propagation time (plus some fixed delays) between the aircraft and the beacon.

Since many aircraft may be interrogating the same beacon, it becomes necessary for the aircraft to distinguish its own reply pulses from many others. This is done by searching for and utilizing those pulses which are synchronous with one's own interrogation transmissions. Accordingly, the search and tracking unit examines all signals and chooses those which have a fixed time with respect to the transmitted signals. The air-borne modulator is designed with an intentional jitter in the repetition rate, so there is little probability that several aircraft can continue to send pulses at exactly the same rate. After the search and tracking unit locks on these synchronous reply pulses, it continually measures the time that elapses from the time of transmission to the time of reception. This action concludes the "search phase" of the equipment and initiates a "track phase" in which the pulse-repetition rate of the interrogations is greatly reduced. This operation reduces the load on the transponder and permits a larger traffic capacity. The time measurement is converted, by suitable instrumentation, to a direct reading of distance in miles.

58. Frequency and Channeling of DME. As mentioned previously, frequencies in the 200- and 1,000-Mc ranges were considered for distance service. An important factor in the ultimate choice was the band width required for pulse transmissions. Distance measurement demands short pulses for accuracy. The half-power band width in megacycles for a square pulse of t μsec duration is equal to $1/t$. Hence, pulse lengths of 2 μsec require a channel width of the order of 1 Mc with some guard space. In the 200-Mc band, modern techniques permit practical selectivities of the order 50 to 100 kc; hence, this band can accommodate many high-information voice-communication channels.[52] At 1,000 Mc, however, the state of the art is such that channel spacing cannot be much less than 1 Mc. Thus for whatever purpose the channels at 1,000 Mc are applied, the channel spacing must be of the order of 1 Mc to avoid interference. Therefore, in view of the need for conserving radio frequencies, the wide spectrum required for distance measurement is more advantageously obtained at 1,000 Mc. This factor, coupled with others involving propagation characteristics and antenna sizes, led to the international standardization of the band from 960 to 1,215 Mc for navigation use.

In determining the number of channels required, the principal factor is the minimum distance at which stations can be located without interference and still supply over-all area coverage. Modern aircraft, flying at altitudes of 40,000 ft, have a radio horizon of approximately 280 miles. It is impossible, however, to space the DME facilities at 560 miles and still supply signals to low-flying aircraft. Accordingly, stations are normally spaced at intervals of approximately 100 miles and sufficient channels must be provided to avoid interference. The problem is further intensified by the addition of DME service to ILS installations.

Cognizant agencies in the United States originally felt that 50 channels would provide adequate service on a noninterfering basis. In the interim period, it was found that this number of channels was insufficient, and present requirements call for at least 100 distance-measuring operating channels. For distance service an operating channel really consists of two channels, one to carry the interrogation from air to ground, and the other for the reply from ground to air.

Two general methods are available for providing these operating channels, namely, the clear-frequency method and the pulse-multiplex method. In the first method, as in Tacan, a 1-Mc channel spacing is utilized, thereby providing 126 two-way clear-frequency channels.

The pulse-multiplex method[53] is employed by the CAA-DME. With this method, each channel employs a combination of frequency and double-pulse codes to distinguish between channels. The pulse code consists of a pair of pulses of certain spacing. These spacings (or modes) are specified by ICAO as 14, 21, 28, 35, 42, 49, 56, 63, 70, and 77 μsec. The same spacings are used for both ground and air transmissions, except that the order is reversed. Accordingly a widely spaced interrogation mode is coupled with a narrowly spaced reply mode and vice versa. These 10 modes can be

associated with any one of 10 frequencies to provide the required 100 two-way channels. These frequency channels are 2.5 Mc wide and are centered, for air-to-ground interrogations, from 963.5 to 988.0 Mc and, for ground-to-air replies, from 1,188.5 to 1,211.0 Mc. As mentioned previously, the ACC decision of 1956 has brought about the decommissioning of the pulse-multiplex DME, and the CAA has scheduled implementation of co-located VOR and Tacan equipments which constitute VORTAC.

59. DME Apparatus. Early DME equipments with relatively wide channel spacings used cavity-stabilized oscillators as the r-f source. Crystal-reference control was introduced later. All modern DME equipments, however, utilize direct crystal control in both the ground and air-borne components. The technique[54] involves frequency multiplication from a crystal-controlled oscillator operating at a subharmonic of the desired output frequency, and lends itself readily to the use of standard vacuum tubes. Peak power outputs of from 6 to 20 kw have been so obtained with a frequency stability of better than 0.01 per cent.

The transponder antenna must be capable of transmitting in the frequency range from 1,188 to 1,211 Mc and simultaneously receive interrogations from 963 to 988 Mc. This virtually means that the antenna must operate over the complete range from 960 to 1,215 Mc. Other requirements of the antenna include vertical polarization, a circular radiation pattern in the horizontal plane, and a power gain in the vertical plane of the order of 8. It is also desirable that the vertical pattern be pointed slightly upward from the horizontal.

One such antenna employs a number of *discone* elements. A discone consists of a metallic cone having an apex angle of about 60 deg and a metallic disk located adjacent to the apex. When fed by coaxial line, the inner conductor may be connected to the cone and the outer conductor to the disk. The discone behaves essentially like a vertical dipole, except that its impedance stays fixed over a wider frequency range.

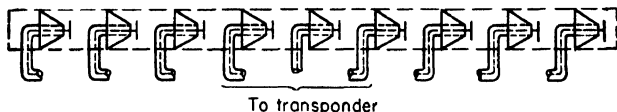



FIG. 34. Discone ground-transponder antenna.

To obtain a power gain of 8 and a beam that is tilted slightly upward, discone elements are stacked as shown in Fig. 34. The individual transmission lines from each discone are connected in parallel at a common junction box, and by means of a $\lambda/4$ transformer, the impedance is rematched to that of the single coaxial transmission line from the transmitter. Mechanically, the individual antenna elements are mounted in a Fiberglas supporting structure, and the entire array is housed in a weatherproof Fiberglas enclosure.

Other essential components of a transponder include a pulse modulator and receiver. The main function of the modulator is to process the interrogation pulses from the receiver so as to apply the desired double-pulse spacing in accordance with the mode selected. A nickel delay line, on which a number of taps or pickup coils are provided, is utilized for the generation of the pulse codes. The input pulse is connected to a pickup coil located at the extreme end of the line and is simultaneously connected to the proper coil for producing the delay defining the particular channel on which it is desired to operate. The pair of pulses so produced are regenerated so as to restore the desired pulse width, and the pulse-pair output triggers the r-f transmitter. The receiver used for the air-borne interrogator and the ground transponder are quite similar in principle, and the following description applies to both.

At the input to the receiver, a series of cavities are used for r-f preselection. In the multiplex-channel system, the separation between the transmitting and receiving frequencies is at least 200 Mc. Hence the former is readily prevented from reaching the receiver and all undesired responses are attenuated by at least 60 db. The remaining r-f portions of the receiver are conventional in nature. The i-f band width is normally 6 Mc, with a channel spacing of 2.5 Mc. To reject adjacent channel transmissions, use is made of a circuit normally referred to as a *Ferris discriminator.* The action of this circuit is shown in Fig. 35. The i-f amplifier has been designed so that adjacent channels are passed

without too much attenuation. Figure 35*A* shows the response of the discriminator as a function of frequency. Similarly, Fig. 35*C* shows the response resulting from a pulse on an adjacent channel. The video circuits following the Ferris discriminator respond only to a negative input. A strong negative pulse occurs only when the pulse spectrum lies centered in the Ferris pass band as shown in Fig. 35*E*. The Ferris discriminator is also useful in the suppression of ignition noise and acts somewhat as a spike suppressor.

The a-g-c requirements are quite different for the transponder as compared with the interrogator. The air-borne receiver attempts to receive emissions from a single transmitter whereas the transponder must be capable of receiving many signals originating anywhere within the 100-mile range of the beacon. In addition, the ground-receiver gain must be controlled to prevent overloading in the ground transmitter.

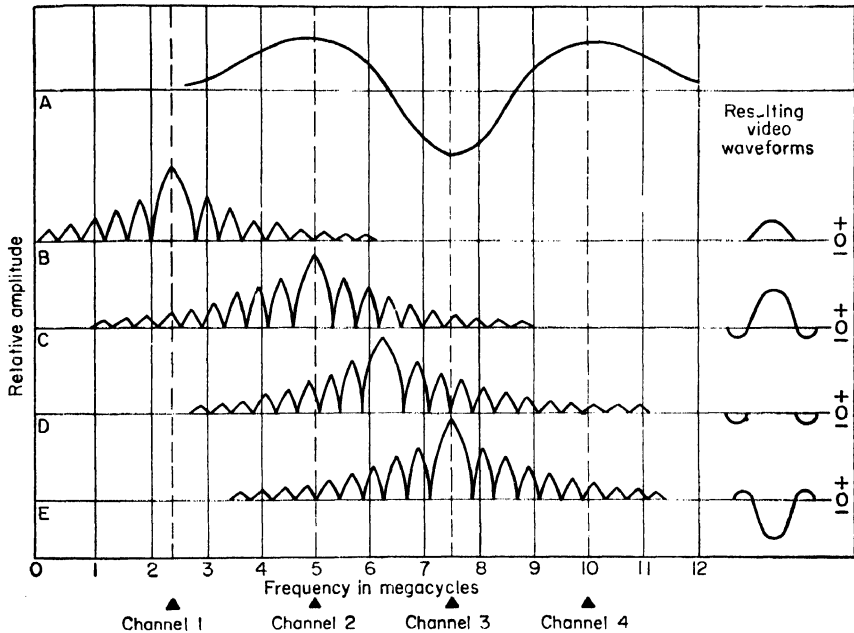



FIG. 35. (*A*) Ferris discriminator response. (*B*) For pulses two channels below the tuned channel, the video wave form is a positive low-amplitude pulse. (*C*) For pulses one channel from tuned channel, the video output is a large positive pulse and two small negative pulses. (*D*) With pulses one-half channel off, the video output consists of two small negative pulses. (*E*) Pulses on the desired channel produce a video output consisting predominantly of a large negative pulse which is passed by the following video stage.

In the case of an interrogator operating on a clear channel, that is, without pulse multiplex, the problem is relatively simple. The AGC is designed to allow normal receiver operation at high sensitivity with a relatively short recovery time. The AGC is controlled solely by those pulses which constitute replies to the particular interrogator. In the case of pulse-multiplex channeling, it is only after decoding that spurious replies are segregated. For the ground receiver utilizing the pulse-multiplex channeling, the receiver is set to have a high sensitivity and designed for rapid recovery. Some of the noise from the receiver is rectified to form approximately 300 pps, and these are supplied to the transmitter. The output of the transmitter is monitored and when the number of pulses reaches 2,000 pairs the gain of the receiver is reduced so that the transmitter cannot be further loaded. This type of control is referred to as *automatic gain stabilization* and *automatic overload control.* A third feature incorporated in the ground receiver is called *echo suppression.* It often happens that interrogations are received by a direct path as well as by reflection from nearby objects. The latter pulses are called echoes and combine with the main pulse to produce a widened pulse. The echo-suppression circuit serves to reduce the gain of the

receiver immediately after the reception of a pulse, but restores the gain of the circuit to normal shortly thereafter.

The principal measuring component of the air-borne receiver is the ranging circuit, or *strobe*. A number of automatic ranging circuits have been employed using a combination of electronic and mechanical techniques. In general, a strobe utilizes delayed gates which are generated by the original interrogations and searched in time for reply pulses. The proper reply signal is identified as a series of pulses coincident in time with the gate. Because of the intentional instability incorporated in the interrogation rates, unwanted replies would not be synchronous with the gate and would therefore appear only occasionally.

Four operations are provided in air-borne DME that automatically and continuously display distance information. These are as follows:

1. *Search.* In this operation the strobe gate is searched in time until the locally generated interrogation pulse and the received pulse occur at the same time.
2. *Lock.* When a delay time has been found that causes the above-mentioned coincidence, and if this condition persists for about a second, the equipment is made ready to enter the track mode.
3. *Track.* During tracking, the delay is shifted in synchronism with the change of distance between the interrogator and transponder so that distance information is continuously and automatically displayed.
4. *Memory.* If the signal is lost momentarily, the equipment enters the memory phase. Two memory systems may be used, velocity memory or position memory. In velocity memory, the indication is rotated in the same direction and at the same velocity as existed prior to the signal loss. In position memory, the indication remains fixed. If the signal is not recovered in a period of approximately 20 sec, the equipment goes into search operation.

60. Range of Accuracy of DME. Range is dependent upon line-of-sight considerations, the power of the transmitter, the sensitivity of the receiver, and the gains utilized in both the transmitting and receiving antennas. Because of the limitations of available power and weight on the aircraft and the necessity for small antenna sizes, the air-to-ground link generally fails first. From a purely theoretical standpoint, a free-space range of approximately 600 miles is possible with a peak power of 1 kw. The transmissions are limited, however, by normal radio line-of-sight considerations, and the useful range of the civil DME, as used by commercial aircraft, can be considered to be approximately 100 nautical miles.

The factors that produce distance-measuring error can be classified as fundamental or instrumentation. Under fundamental, the most important is the pulse-rise time. With the ICAO system, a pulse-rise time of $\frac{1}{4}$ μsec corresponds to an error of about 0.02 nautical mile. The rise time is of course limited by the allowable band width. A second fundamental error arises from bending or ducting in the atmosphere, but is not a serious factor at 1,000 Mc. A third fundamental error is the variation in the speed of radio propagation.[55] Finally, the large source of error, in so far as navigation is concerned, is that DME measures slant range rather than the geodesic. At high altitudes and close to the station, this error may be very large and, in fact, becomes an altitude indication over the beacon.

The instrumental errors are mainly attributable to the instability of components and the inaccuracies in the generation of frequencies and delays used in the measuring technique. Actual operational errors with existing equipment can be expected to be between 0.1 and 0.25 nautical mile.

61. Arbitrary-course Computers.[56,57] The addition of DME to VOR constitutes a polar-coordinate system where the origin of the coordinates is the location of the station. This type of presentation has the advantage of being the traditional system of geographers and lends itself readily to mechanical position plotting. In addition, this presentation permits the simple computation of arbitrary courses as illustrated in Fig. 36. In the illustration, an aircraft is located at point A at a distance R from a beacon located at S. The bearing of the aircraft from the station (referred to north) is θ. The aircraft desires to fly to an arbitrary destination on a course having a bearing

of ϕ with respect to north. The perpendicular distance from the course to the radio station is D. At all times, the equipment in the aircraft produces indications of R and θ. The arbitrary course is completely described by D and ϕ; therefore the computer is required to solve all values of R and θ that lie on the course. Relatively simple mechanical computers have been developed to perform this function. The results of the computation are presented on the cross-pointer instrument, and the arbitrary course is flown by keeping the instrument centered.

62. Some Avigational Aids for the Short-distance En-route Zone. In addition to VOR and DME, a number of other aids have been employed successfully in short-distance en-route zones. Although not considered for international standardization, these aids serve a useful purpose and, in some cases, are better suited to the specific avigational tasks in those particular areas. This section briefly describes three of

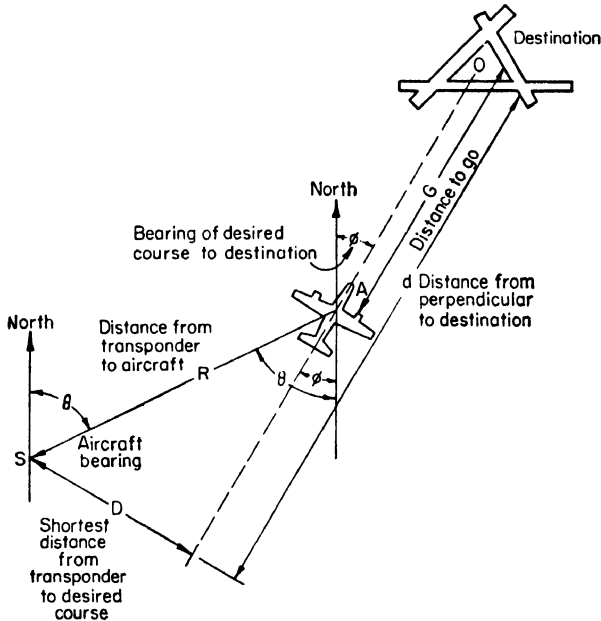



FIG. 36. Diagram of arbitrary-course computation.

these aids; Gee, developed by the British; the visual-aural range developed by the United States CAA; and the v-h-f DF introduced during the Second World War.

During the postwar period numerous new systems for navigation and traffic control were proposed, and a substantial number of these have received some developmental effort. Space does not permit a description of all of these, but a few have been chosen which have undergone considerable development and which incorporate principles that are likely to be used in future developments.

63. Gee. The Gee system was first proposed by R. J. Dippy[58,59] of England in 1937, completed development in approximately 1940, and was used by both British and United States aircraft operating in the European Theater during the war. Although the system was considered, but not accepted, by the Special Radio Technical Division of ICAO, Gee chains have continued to operate in Europe.

Gee is a true hyperbolic system employing synchronous transmission of pulses from three or four ground stations in exactly the same manner as employed by loran. Gee, however, operates at the higher frequencies of 22 to 30 and 40 to 85 Mc. Because the frequencies are quasi-optical, the distance between pairs of a Gee station is 70 to 100 miles. The

pulses transmitted have durations varying from 2 to 10 μsec. The master station of the chain employs a crystal-controlled repetition rate of 500 main pps; however, every other pulse is followed by an auxiliary pulse so that the actual number transmitted is 750 per sec. The second pulse serves as means for identifying the various signals as they appear on the display. The sequence of pulse transmissions is shown in Fig. 37. The first slave station *B* receives the master pulses and after a fixed delay transmits only single pulses at a pulse-repetition rate of 250. The second slave *C* operates in a similar manner and thus incorporates a pulse-repetition rate of 250. The third slave or *D* station transmits a double

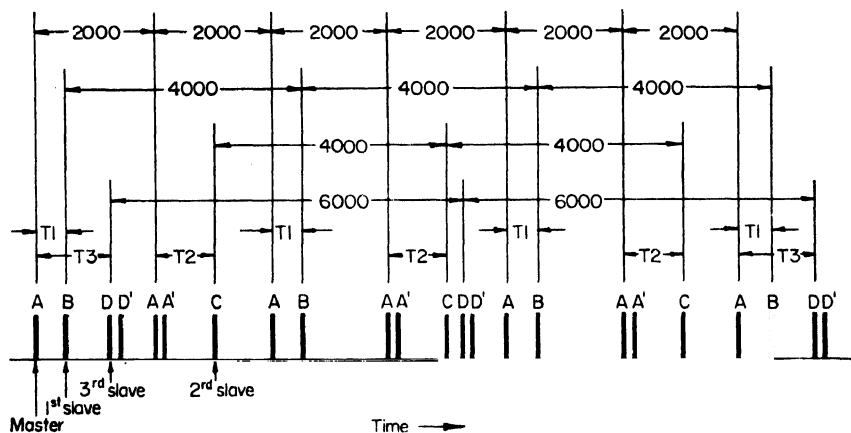



FIG. 37. Pulse sequence in a four-station Gee system.

pulse after receiving a single pulse from the master station but has a pulse-repetition frequency of $166\frac{2}{3}$. It can be seen from Fig. 37 that, although the positions of the *B*, *C*, and *D* pulses will change with position of the aircraft, it will always be possible to identify which pulse arrives from which station.

The Gee ground equipment employs techniques which closely parallel those used for loran. Originally, the pulse receivers at the slave stations were connected directly to the Gee transmitter. It was found, however, that fading over the path of propagation between the master and the slave station was sometimes excessive and special circuitry was adopted to overcome this problem. This circuitry involves a stable 150-kc crystal-controlled oscillator at the slave station which is controlled in frequency and phase by the incoming master pulses. The incoming pulses are used to shock-excite a 150-kc crystal filter whose output is compared in a comparator circuit with the output of the crystal-controlled oscillator. Any errors in frequency or phase are used to control a servoamplifier which corrects the crystal oscillator. The oscillator frequency then triggers the transmitter in parallel with the output from the receiver. Accordingly, if the output of the receiver fails from time to time, the required pulses are supplied by the crystal oscillator with no interruption in the transmission.

FIG. 38. Gee display with calibration marks at 150 and 15 kc.

In contrast to loran, stations are not selected by their differences in pulse-repetition rate, but solely by choosing the radio frequency associated with the desired chain.

A typical Gee presentation is shown in Figs. 38 and 39. The measurement technique consists in first identifying the various pulses and then measuring the required time intervals between pulses using accurately controlled time marks. After initial identification using a main time base, a switch is manipulated to change to a strobe time base. The strobe time base is much more rapid and, by adjustment, permits the leading edges of the pulses to be closely aligned. The next step involves the measurement of time difference which utilizes three accurate frequency sources of 150, 15, and 3 kc, respectively. The display then appears as shown in Fig. 38. In this display the marks generated by the 15-kc frequency appear taller than those generated by the 150-kc frequency. The last significant figures of time delay are determined by counting the number of marks between the large mark on the center of the *A* pulse sweep and the large mark on the left of center on the *B* pulse sweep. It is also necessary to estimate the fraction of the scale division to obtain the last significant figure. Next, the main time base (slow sweep) is used to replace the fast sweep and the display than appears as shown in Fig. 39. In this display, the 15- and the 3-kc time sources are used, and by a process similar to that described it is possible to count the divisions corresponding to the first two significant figures of the time delay. In the examples shown in Figs. 38 and 39 the reading is 11.66.

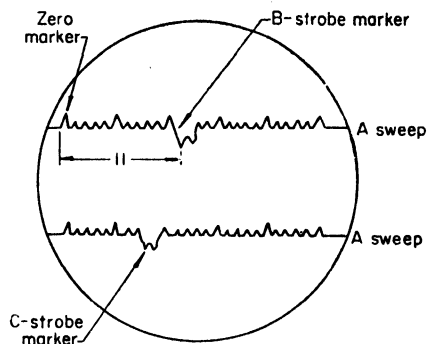



FIG. 39. Gee display with calibration marks of 15 and 3 kc.

The coverage of the Gee system is dependent upon whether two or three slave stations are used. With two slave stations, the area of best performance is limited to approximately 120 deg. To obtain 360-deg coverage, it is necessary to employ a master and three slave stations. At an altitude of 15,000 ft, performance can be expected to a distance of approximately 400 miles from the master station, but at an altitude of 5,000 ft, the range is limited to approximately 300 miles. Assuming that the time difference can be measured with an accuracy of approximately half the distance between the 15-kc marks, an error of approximately ½ μsec results corresponding to a distance of 150 ft. This accuracy can be expected where signal strength is high and the distance is not greater than the length of the base line (70 to 100 miles). At greater distances, accuracy of position fixing is considered to be about ½ per cent of the distance. At maximum range, the error distribution becomes long and narrow and the information received is applicable to determining a line of position rather than a fix.

64. The Visual-Aural Two-course Radio Range (VAR). This system was developed[60] by the CAA beginning in 1939 and constituted an interim navigational aid on the United States civil airways. The device was intended to replace the four-course m-f radio range. In 1941, eight VARs were installed on the Denver–Los Angeles airway. Following the Second World War, the New York-Charleston airway was similarly implemented,[61] but after 1946 it was decided that the omnidirectional radio range was required to provide multiple airways; so no more installations were made. It should be noted, however, that as late as 1951 the two above-mentioned airways were still in operation and provided satisfactory service. Operational experience indicates that the VAR is a satisfactory device where only a single airway is desired.

VAR operates in the 100-Mc band. The space patterns utilized by the device are shown in Fig. 40. These patterns consist of two intersecting lobes, one modulated with 90 cps and the other with 150 cps. The course information is derived by detection and comparison of amplitude of these tones. The relative amplitude is displayed on a zero-center meter which may deflect right or left depending upon which tone is greater in magni-

tude. The course lies at the intersection of the patterns where both tones are equal. This constitutes the visual portion of the range.

In addition to the pair of patterns modulated with 90 and 150 cps, a second, similar pair of patterns is modulated with 1,020 cps keyed in an interlocking Morse-code fashion to produce the characters A and N. The principle employed here is the same as that used for the four-course radio range previously described. Accordingly, an aural course is formed at right angles to the visual course, and is identified as a steady 1,020-cps note. In general, the pilot flies the aircraft by visual reference to the right-left meter and utilizes the aural course only to orient himself with respect to the radio station. In effect, the system provides quadrant identification.

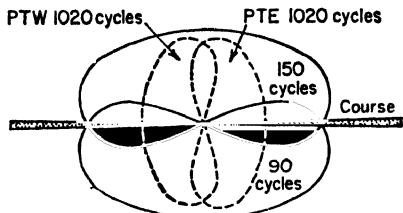



FIG. 40. V-h-f visual two-course range with aural-sector identification.

The transmitting equipment employs techniques which are similar to the localizer of the fixed-beam low-approach system. These techniques are discussed in Part 3. The receiver used with the visual-aural range is the same as that required for the localizer and is also described in Part 3. Briefly, it consists of a crystal-controlled superheterodyne with good AGC. Filters are used to separate the 90- and 150-cps components, which are then compared in magnitude in a bridge-rectifying circuit. The resulting d.c. is applied to a zero-center meter to indicate course location.

The range of the VAR is governed by line-of-sight considerations and can be expected to be approximately 100 miles for commercial aircraft. By flight experience, the optimum course width has been found to be 20 deg, defined as a full-right to full-left deflection.

65. V-h-f Ground Direction Finders. During the Second World War, direction finders (DF) taking bearings on v-h-f transmissions became the primary means of navigation for fighter aircraft. ICAO, however, would not consider these devices as standard aids because of the limited traffic-handling capacity. Nevertheless, their usefulness was realized and, in response to requests from several of its regions, ICAO developed a standardized procedure for use of ground direction finders.[62]

In 1948, Special Committee 31 of the RTCA, which was studying the problem of air-traffic control, recognized the value of v-h-f ground DF and recommended[63] the installation of these units wherever surveillance radar was used. In this application, the DF provides a means of identifying targets on the radar scope. In many areas of the world where traffic densities do not saturate ground direction-finding operations, the device is capable of providing guidance to aircraft which are equipped with a minimum of radio apparatus. In fact, to obtain this service, the only air-borne equipment required is the usual two-way v-h-f communication set.

The principles of operation are discussed in Sec. 31. The antenna system almost universally used for direction finding at very high frequencies is the elevated H Adcock, which has the advantage of low polarization error and large effective height.

The accuracy of a v-h-f DF depends upon the combined effects of instrument errors, polarization errors, and site errors. The inherent instrumental accuracy of automatic DF is kept to an order of about 1.5 deg. Polarization errors on a standard wave may be kept within approximately 4 deg. Site errors are difficult to estimate and vary widely with azimuth. In general, direction finders can be said to have an over-all rms accuracy of better than 6 deg, and under conditions of good siting and accurate calibration, accuracies considerably better than this are possible.

66. The Coordinated-system Concept. The concept of a coordinated system was advanced following the Second World War, particularly for aerial navigation and traffic control, and was accepted as the principle for the ultimate aerial navigation and traffic-control system by Special Committee 31 of the RTCA.[63] The concept provides economies in both the radio spectrum and air-borne equipment. It has been shown by H. Busignies[64] and others that the amount of information required by aerial-navigation systems is small, and the coordinated system concept can be stated briefly

as: "Since the amount of information required to be transmitted for each aerial navigation and traffic function (that is, distance measurement, bearing, instrument approach, etc.) is small, then the most efficient system is one in which all information is carried on a single r-f channel."

The concept is applied practically in the following manner. Every navigation device can be considered to consist of two main parts. The first part is concerned with generating or utilizing an r-f channel, and the second part either impresses navigational information on the channel or extracts it and presents it to a suitable indicator. In conventional systems that have been employed in the past years, each function makes use of a separate group of channels; therefore, it was necessary to employ a separate r-f section (whether it be transmitter or receiver) for each function. The coordinated system can be imagined to be a single transmitting or receiving equipment and all translators, regardless of function, are associated with these two r-f sections.

Shannon[65,66] and others have developed a theory for the information capacity of a channel. The capacity in binary digits, or bits, is given by the expression $C = W \log_2 [(P + N)/N]$.

In this formula, W is the width of the channel in cycles, P is the power of the signal, and N is the power of the noise expressed in the same units as P. For a channel 1 Mc wide and a signal power twice that of the noise power the channel capacity is 1,590,000 bits. White has made the following estimates for the devices employed in short-range navigation, landing, and traffic control:

1. Distance-measuring service: 3,000 bits
2. Beacon identity and altitude: 2,000 bits
3. Azimuth: 48 bits
4. Localizer: 306 bits
5. Glide scope: 306 bits
6. Marker: 50 bits

The total of the above is 5,710 bits. It cannot be presumed that a channel can be used to its maximum theoretical capacity. But since the ratio between the amount of information required and the capacity of the channel is about 28:1, it seems reasonable to assume that a practical system could be evolved.

Another important consideration, however, is the reliability of the coordinated system. It can be assumed that the electron tube is the chief cause of air-borne radio-equipment failure. Assuming a 10-hr flight, the probability of failure of various equipments has been estimated as follows:

1. Distance-measuring equipment: 21 per cent
2. Beacon: 15 per cent
3. V-h-f omnidirectional range receiver: 9 per cent
4. Localizer receiver: 6 per cent
5. Glide-scope receiver: 6 per cent
6. Marker receiver: 3 per cent

If, however, the complete system is important to the fulfillment of a mission, then the reliability of the complete system must be considered rather than the reliability of each individual unit. The total number of electron tubes (neglecting power supply) of the present conventional equipment listed above is 151. In the coordinated system, it is estimated that this number could be reduced to 76. For the conventional system the aggregate probability of failure, for a 10-hr mission, is 45 per cent. For the coordinated system, because of the reduction in the number of tubes, the equivalent probability is 26 per cent. It is thus seen that the reliability of the coordinated system is greater than that of the conventional system taken as a whole but is less reliable than any of the individual units of conventional design. However, because of the savings in the number of tubes of the coordinated system, it is possible to duplicate the system. With duplicate coordinated equipment, the probability of failure is reduced to only 7 per cent.

Accordingly, for approximately the same number of tubes, two coordinated air-borne units may be employed, thus reducing the probability of failure far below that of the conventional system as a whole, and approximately equal to the most reliable component of the conventional system.

The above constitute the main arguments of observers favoring the adoption of a coordinated system of navigation. The following sections describe three systems which utilize these principles at least in part and comprise the Navar Coordinated system, the Sperry Microwave Coordinated system, and the SC-31 system. It should also be noted that Tacan was conceived as a coordinated system, and the compatible addition of ILS service and two-way data transmission has been demonstrated.

67. Navar Coordinated System. Navar is the name coined by the Federal Telecommunication Laboratories for a system of aerial navigation and traffic control on which development started in 1945.[67] As shown in Fig. 41, the Navar system made use of two sets of receivers and transmitters on the ground operating at 3,000 and 1,000 Mc, respectively. The air-borne equipment employs a 3,000-Mc crystal video receiver and a 1,000-Mc transmitter and receiver. With this equipment, the aircraft is furnished with distance measurement, bearing, relayed radar, and traffic-control advice.

On the ground, the 3,000-Mc transmitter and receiver furnish conventional surveillance radar service, and the display incorporates identity-altitude information in the form of numbers. The distance service is furnished in a conventional manner utilizing the 1,000-Mc ground transponder and the 1,000-Mc air-borne interrogator. The details of operation are similar to those already described in Sec. 57 (distance-measuring equipment).

Bearing service is obtained on the 3,000-Mc equipment which utilizes a highly directive rotating beam on the ground. The beam width used is 3 deg. Without interrogation, the 1,000-Mc transmitter on the ground automatically and continuously sends out a distinctively coded set of pulses via its own antenna each time the 3,000-Mc rotating antenna is pointing northward. Since the 1,000-Mc transmission is omnidirectional, all aircraft receive these reference pulses simultaneously. After decoding, the 1,000-Mc reference pulse is used to operate a timing circuit for bearing measurement. The time interval between reception of the reference pulse and the 3,000-Mc radar beam is a function of bearing, this quantity being displayed on an indicator calibrated in degrees from north.

The radar pulse received on the 3,000-Mc receiver serves to trigger the 1,000-Mc transmitter and hence constitutes a beacon reply to the ground. This beacon reply is received on the ground 1,000-Mc receiver and displayed on a conventional radar display except that there are no permanent echoes.

The north pulse received by the 1,000-Mc receiver in the aircraft is used to synchronize the rotation of a deflecting coil on a cathode-ray tube in the aircraft. On the ground, the video output of the radar is used to modulate the 1,000-Mc transmitter. The signals received on the 1,000-Mc air-borne receiver, when applied to the cathode-ray tube, therefore reproduce the ground radar display.

The 1,000-Mc transmissions from the aircraft also convey information to the ground as to the altitude and identity of the aircraft. The altitude is encoded as the spacing between a pair of pulses as governed by an altimeter. The identity is relayed by a special group of pulses arranged in a certain manner by a coding unit. To determine the identity and altitude of an aircraft, it is necessary to operate decoder controls that set a radial and circular cross hair to intersect on the desired target on the beacon-reply display on the ground. This operation decodes the reply, which then appears as a series of numbers on a separate display. With the decoder controls adjusted in the manner just described, it is also possible to send a short discrete message to the aircraft. The elements of the Navar system are shown in Fig. 41.

The Navar system was developed to the point where it was demonstrated as a working unit and was used to prove the practicability of the coordinated-system concept. Many of the techniques used in Navar formed the basis of the Tacan development which followed.

68. Microwave Coordinated System. The Sperry Gyroscope Company has engaged in the development of a coordinated system which is designed to operate at 5,000 Mc and provides bearing-distance service. Because of the difficulty in obtaining very narrow

selectivity in the 5,000-Mc range, the required number of operating channels for a territory with dense air traffic is secured by the use of subcarrier modulations, and the entire system is operated on a c-w basis. With this operation, various services can readily be multiplexed on the same channel by using a different kind of modulation for each service.

The block diagrams of the system are given in Figs. 42 and 43 for the air-borne and ground equipments, respectively. In the aircraft, the distance measurement is accomplished through the use of two modulating frequencies, namely, 1,000 and 18,600 cps. The signal is received by a microwave receiver on the ground and retransmitted on a different frequency by the transmitter, which is also employed in bearing service. Distance is

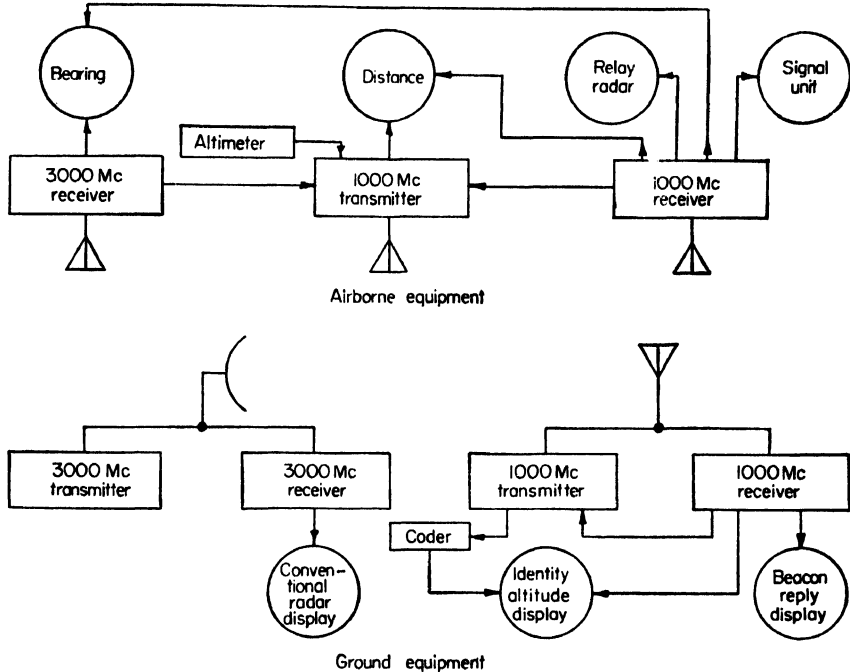



FIG. 41. Block diagram of IT&T Navar coordinated system.

derived by phase comparison in the aircraft of the two modulating tones. For example, at a distance of 93 miles, the 1,000-cps transmission relayed by the ground returns to the aircraft at 360 deg out of phase with the transmitted wave. At a distance of 5 miles from the station, the 18,600-cps tone undergoes a similar phase shift of 360 deg. The phase shift of the 18,600-cps tone will, in fact, reproduce every 5 miles and therefore requires the 1,000-cps tone to resolve the ambiguity. The phase detectors associated with the 1,000- and the 18,600-cps frequencies are combined in a selector circuit which controls the rotation of two phase shifters that have suitable gearing between them. This mechanism automatically resolves the ambiguity and displays the information in miles.

The mechanism employed to produce bearing is designed for high accuracy and differs radically from that employed in the VOR. On the ground, a rotating pattern is employed which produces 11 lobes instead of a single rotating limacon. These lobes are received as a sinusoidal modulation in the aircraft receiver and will undergo a phase shift of 360 electrical degrees for approximately every 33 deg of bearing. This vernier effect can theoretically increase the bearing accuracy by a factor of 11; however, there are 11 points of ambiguity that must be resolved. This is accomplished by combining the 11-lobe pattern with a single limacon pattern, the latter selecting which of the eleven 33-deg sectors is the proper one.

The ground antenna is rotated at a speed of 1,650 rpm, which produces a modulation of

27.5 cps from the limacon and a frequency of 302.5 cps from the 11-lobed pattern. A voltage of 27.5 cps is generated by a permanent-magnet alternator associated with the antenna shaft, and this frequency is used to constitute the reference signal. The reference is impressed on the carrier by a combination of frequency-modulation and subcarrier techniques.

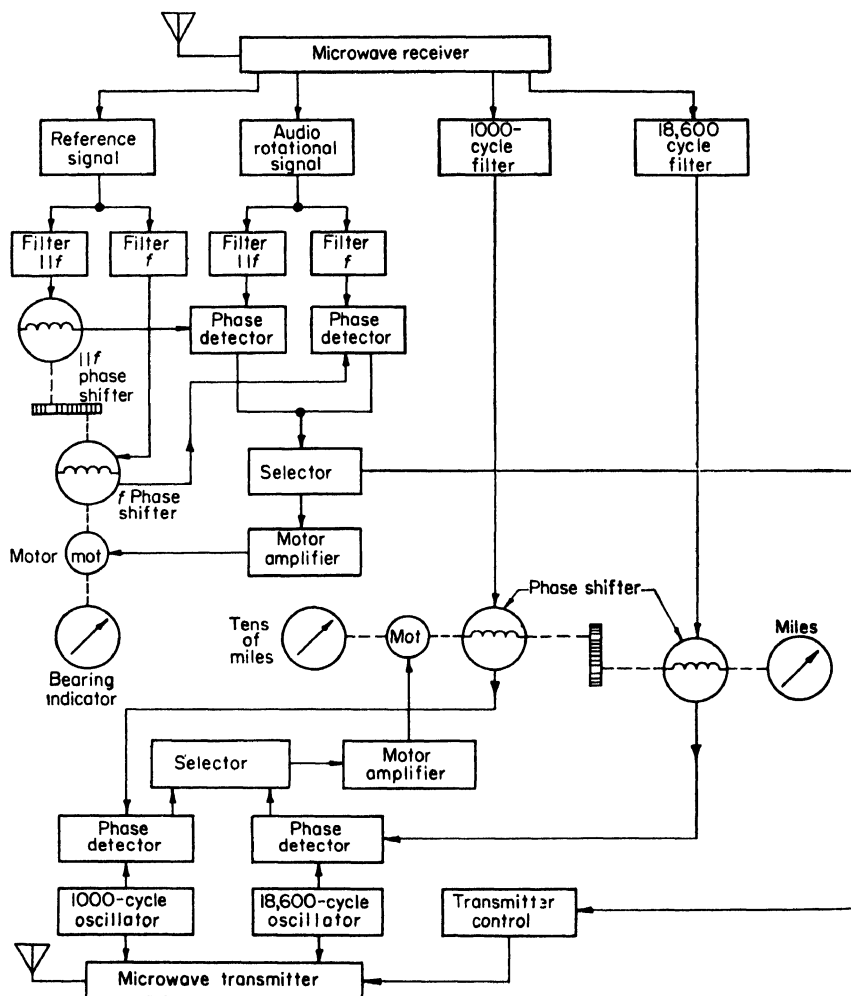



FIG. 42. Block diagram of the air-borne equipment of the Sperry microwave coordinated system.

Azimuth and distance scanning is possible with this system by an arrangement that affords complete position information on the ground corresponding to all aircraft within the area. A separate microwave link is required for this service. The ground transmitter of the link is modulated by a signal which is derived from the reference signal of the omnirange but is caused to vary continuously in phase at a constant rate of, say, 60 deg per sec by passing the omnirange reference signal through a continuously varying phase shifter. This signal is referred to as the variable-phase link signal. On the aircraft, there is already available the variable-phase signal from the omnirange, the phase of which is a direct function of the bearing of the aircraft. A second phase-comparison circuit is provided

on the aircraft which compares the variable-phase signal from the omnirange receiver with the variable-phase link signal.

The aircraft transmitter for the separate microwave link is normally in stand-by condition. Its modulation consists of a frequency corresponding to the distance from the ground station and developed in the aircraft as a result of the distance-measuring equipment. As the variable-phase link signal from the ground station passes through phase synchronism with the variable omnirange signal at the aircraft, the output of the second phase-comparison circuit is made to trigger the transmitter. The ground-station receiver then obtains a signal which indicates the distance of the aircraft. The arrangement described also comprises an azimuth-scanning system. Since the signaling phase shifter is varied at a constant rate, the transmissions from aircraft at different azimuth bearings will be separated from each other on a time-sequence basis. The distances of the aircraft will be distinguished from each other on a modulating frequency basis. If two or more aircraft appear at the same azimuth bearing, their transmitters will respond at the same time but will show different modulations according to their distances.

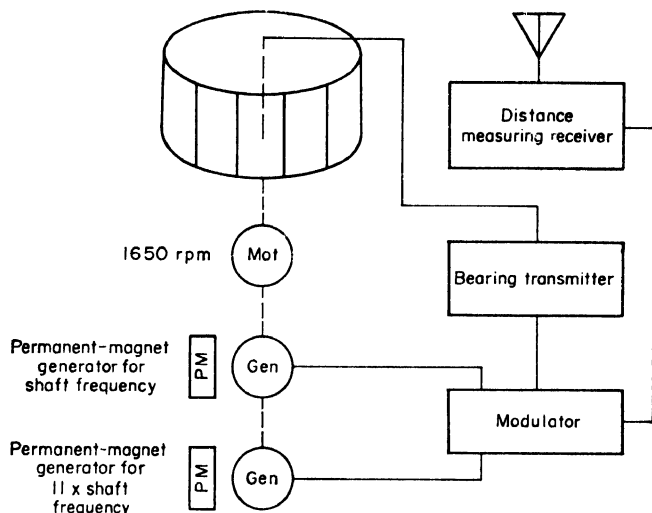



FIG. 43. Block diagram of the ground equipment of the Sperry microwave coordinated system.

The system also permits retransmitting the ground-station information back to the aircraft, thus providing information as to the bearing and distance with respect to the ground station of all aircraft in the area.

A model of the microwave system was constructed and tested to prove the soundness of the principles involved.

69. The SC-31 Coordinated System. Special Committee 31 of RTCA recommended[63] that the coordinated-system principle be applied to a number of equipments in the ultimate system intended to replace the interim common system of aerial navigation and traffic control. Specifically, the committee recommended that the navigational equipment be a transmitter-receiver having multiple channels capable of providing the following services:

1. Distance and bearing information for navigation. These data, when used in conjunction with a computer, should allow the pilot to fly any desired course.
2. Precise slope, localizer, and distance information for instrument approaches.
3. Information for airport-surface navigation to enable the pilot to taxi the aircraft.
4. Air-ground aural communication of a reliable and static-free type.
5. A suitable display in pictorial form that enables the pilot to monitor traffic conditions in his vicinity or receive other pertinent data such as holding areas, air-line locations, and weather maps from the ground.

6. Suitable output to allow the aircraft to be flown automatically either en route or during final approach and landing.

The present state of the art indicates that these ultimate recommendations are attainable, at least to a considerable degree.

70. Tacan. Tactical requirements of the military services prompted the development of the Tacan (tactical air navigation) system. Its background includes the early work on 1,000-Mc equipment (such as Navar) from which its techniques were drawn. These early equipments also included IFF radar and distance-measurement equipment for the armed forces. The main contribution of this early work was the development of 1,000-Mc crystal-control techniques and pulse circuitry which formed the foundation of later equipment.

In 1947, the Navy determined the need to replace its YE/YG beacon, an aural v-h-f navigation system. The Navy undertook tests of existing civil-air-navigation equipment of that time to determine whether any met the requirements, but found that in shipboard use bearing errors were excessive. Accordingly, in 1948, a contract was issued to the Federal Telecommunications Laboratories for the development of a 1,000-Mc omnibearing system suitable for shipboard use. This equipment, designated AN/URN-1 (ground transmitter) and AN/ARN-16 (air-borne receiver), utilized a three-lobe system for bearing information and was suitable for integration with the USAF AN/APN-34 DME. In 1949 a second contract was awarded for integration of the two systems resulting in the AN/URN-3 (XN-1) ground equipment and the AN/ARN-21 (XN-1) air-borne equipment. Contracts were awarded for service-test models of the AN/URN-3 (XN-1) and AN/ARN-21 (XN-1) in 1950. However, these contracts were modified in 1951 to provide service-test models of the equipments to satisfy the joint Air Force and Navy requirements. The main differences between early test models and subsequent production models of AN/URN-3 and AN/ARN-21 equipments were the increase in the number of channels from 50 to 126, the improvement of bearing accuracy by substitution of a nine-lobe system for the three-lobe system, and the increase in range capability to 200 nautical miles.

Fig. 44. Tacan instrumentation providing bearing and distance information from a known reference point.

The Tacan equipment is presently being produced in large quantities for the military forces and has been installed for operational use on many carriers, shore stations, and military air bases. Consideration was given to adoption of Tacan for civil use, and in the fall of 1956, the Air Coordinating Committee recommended the conversion of civil DME to the Tacan channeling. This type of DME, sometimes referred to as DMET, when combined with the civil VOR for bearing, has been termed VORTAC. With VORTAC, any Tacan air-borne equipment can be utilized for civil DME service.

Tacan is a radio aerial-navigation system of the polar-coordinate type illustrated in Fig. 44. That is, there is a *bearing* facility that provides the aircraft with a meter indication of its direction in degrees of bearing from the ground beacon selected by the pilot. Also, there is a *distance* facility that provides the aircraft with a meter indication in nautical miles of its distance from the ground beacon. Knowing bearing and distance from a specific geographic point, the pilot can fix his position on a chart.

71. Principles of Tacan. The radio elements of Tacan are indicated in Fig. 45. The entire system operates in the 960- to 1,215-Mc band. One multichannel air-borne receiver-transmitter, operating with pulses, provides both distance and bearing functions. That is, the same radio signals transmitted over a selected channel convey both distance and bearing information. Actually, both a coarse and a fine indication are combined to produce accurate bearing measurements. Only one air-borne antenna and channel selector are required.

Tacan has 126 two-way operating channels of 1-Mc spacing. For air-to-ground transmission (required only for the distance function), there are 126 frequencies within the band 1,025 to 1,150 Mc. For ground-to-air transmission (serving both bearing and distance functions), there are 63 channels between 962 and 1,024 Mc and a like number from 1,151 to 1,213 Mc. The channels are clear frequencies established solely on the basis of r-f selectivity; they do not depend on pulse coding. Double pulse coding is used to increase the effective radiated power and the signal-to-noise ratio.

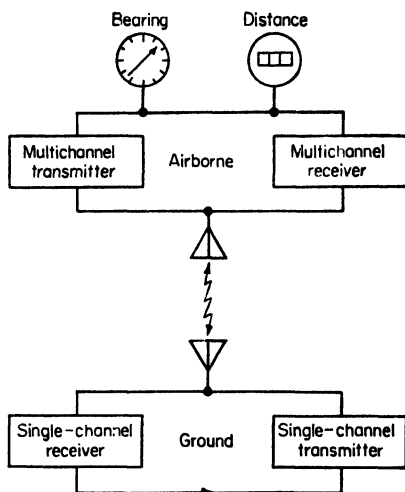
Fig. 45. IT&T Tacan system providing 126 channels for air-to-ground beacon system.

Being conceived as a coordinated system, it is possible to multiplex on the Tacan channels such additional navigational functions as localizer, glide slope, marker, and data transmission. These functions have already been demonstrated experimentally with existing air-borne equipment. When the additional functions are so multiplexed, no further r-f air-borne radio equipment is required. It is only necessary to add adapters and indicators.

72. Tacan Distance Measurement. The principles of distance measurement are identical to those previously described for the civil DME (Sec. 57) but are repeated here to clarify some of the significant differences in channeling and pulse coding.

A given ground beacon may be interrogated simultaneously by a number of aircraft and will reply to all interrogations. Each airplane will receive all the replies being transmitted. To permit interference-free operation under these normal conditions, the interrogation pulses of each airplane occur at a rate that is intentionally permitted to "jitter," or vary within certain limits in an irregular or random manner, by permitting a nonstabilized multivibrator circuit to exercise partial control over the interrogation rate. To sort out those pulses that are replies to its own interrogation pulses, the ranging circuit makes use of an automatic stroboscopic search process.

The strobe locates the reply pulses by finding the one fixed, or very slowly changing, time delay, *always measured from its own previous interrogation pulse,* at which a reply pulse is repeatedly received. Because the interrogation pulses from other aircraft are nonsynchronous or random with respect to those from a given aircraft, reply pulses corresponding to such foreign interrogation pulses will not be received regularly at any fixed time delay on the given aircraft.

The strobe searches automatically each time the air-borne radio set is tuned to a new ground-beacon channel or if there is some *major* interruption in the radio signals. The strobe scans progressively over various time-delay intervals by means of a sliding range gate or time slot. It tests each time-slot position for the number of successive reply pulses received within a certain uniform checking period. If no replies, or only sporadic replies, are received, the strobe advances the range gate to test a slightly longer time-delay interval, and so on. When, at some particular time-delay interval, safe evidence of recurrent replies is detected by a counting process, the strobe has completed its search and stops, since this condition is fulfilled only by reception of the desired reply pulses. Those pulses are the

only ones that are always received in synchronism with the randomly jittered interrogation pulses.

The complete "search" process, covering the entire range of time delays, may require up to a maximum of 20 sec. Thereafter, the strobe locks to the proper reply pulses and converts to a "tracking" operation. Then the delay setting of the range gate automatically and continuously follows any normal variations in the time delay of the proper reply pulses. Such variations occur as the distance between the aircraft and the beacon changes as a result of the flight path.

When the strobe is locked to the proper reply pulses, the time-delay setting of the range gate is a proportionate measure of the distance of the aircraft from the ground beacon (approximately 12 μsec round-trip travel time per nautical mile). The mechanical position of the device that varies the time delay of the range gate is used to control electrically the position of the numerical indicators on the distance meter.

A crystal-controlled 4,044-cps oscillator is used as a time reference. For this frequency, the period $1/f = 242$ μsec, which is the round-trip travel time to an object 20 nautical miles away. In essence, the time interval between interrogation and reply (the time-delay setting of the range gate) is measured in terms of the corresponding number of cycles and fractions of a cycle of the 4,044-cps reference wave.

The time-measuring circuits have a memory provision, so that, if reply signals fail to be received for approximately 10 sec, the existing distance indication will be maintained without the search operation being restarted.

The ground beacon transmits a station-identification signal in International Morse code about every 75 sec and temporarily replaces all reply pulses that would normally occur at that time. The memory feature prevents this substitution from interrupting the air-borne distance indications. On the aircraft, the identification signal is received audibly in telegraphic code.

By strobe principles, which are common to all types of distance-measuring equipment, many (more than 100) aircraft may simultaneously, and without mutual interference, obtain distance service from one ground beacon. Pulse interrogation rates from each aircraft are of the order of 24 per sec during tracking operation; they are temporarily higher, approximately 150 per sec, during the search process. The over-all average rate is considered to be about 30 per sec on the basis that the average aircraft within range will be tracking 95 per cent of the time and searching 5 per cent of the time. The pulses are of the order of 3.2 μsec wide and the duty cycle is very low.

Actually, the pulse signals are always twin pulses, with a prearranged spacing between the two components. The receivers, both ground and air-borne, are followed by twin-pulse decoders that pass only pulse pairs of the prescribed spacing. Isolated single pulses, or pulse pairs with some other spacing, will not pass the decoder. This technique increases the average power radiated and makes the system less susceptible to errors or interference caused by false signals. False signals might be produced by radars, ignition systems, and other extraneous sources, but only rarely will these have the right spacing to pass twin-pulse decoders.

In principle, the transmitter of a distance-measuring ground beacon need not send out pulses except in reply to interrogations actually received. For each such aircraft, it would send out approximately 30 pulses per sec. The total output of the beacon transmitter might then vary from zero to about 3,000 pulses per sec, depending on the existing traffic.

However, one of the major problems with all transponders or beacons is to maintain the receiver sensitivity at the maximum level without having the associated transmitter continually triggered by noise. A very satisfactory solution is to operate the beacon on the constant-duty-cycle principle. In this method of operation, the beacon receiver is provided with an AGC that maintains the number of pulses out of the receiver at an almost constant value, in this case at about 2,700 pulses per sec.

Thus, if few interrogations are being received, the gain of the receiver automatically increases, adding noise-generated "filler" pulses until the requisite total number is obtained. If several more aircraft come within range, the receiver gain drops slightly and the noise-generated filler pulses are replaced by replies to interrogations. If more than 100 aircraft interrogate, replies are made only to the strongest 100 pulses. The relation between gain and number of pulses is such that only a 3-db change in sensitivity occurs between the reception from 1 aircraft and that from 100 aircraft.

An added advantage of this type of operation is the constant current drain on all transmitter circuits. This permits simpler power-supply design, minimizes drifts of tuned circuits, and allows a more predictable component life and maintenance cycle.

Constant-duty-cycle operation has a further advantage. With the beacon always transmitting some 2,700 pulses per sec, there is provided a medium over which other navigational functions can operate. As far as distance information is concerned, it is only the timing of the pulses that matters. Variations in strength of the pulses, produced by pulse-amplitude modulation, may be used to convey additional information. The Tacan bearing function applies this principle.

73. Tacan Bearing Measurement. The heart of any omnidirectional range is the ground-beacon antenna system. It must produce a specified directional radiation pattern that is rotated or switched around a vertical axis. The physical dimensions of an antenna depend on the operating radio frequency or wavelength, and there is a 10:1 reduction in wavelength between the v-h-f band, in which present omnidirectional ranges operate, and the 1,000-Mc region used for Tacan. Advantage is taken of this factor to increase the Tacan-antenna size considerably (in terms of wavelength) for superior performance, while still maintaining reasonably small physical dimensions.

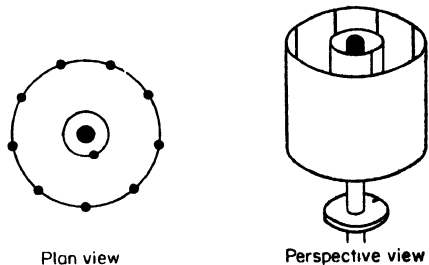
The Tacan ground antenna is fully enclosed in a cylindrically shaped protective cover that is approximately 3½ ft in diameter with a total height, including antenna proper and motor drive, of about 6 ft. No counterpoise is required and the antenna may be installed on a tower or on the mast of a ship.

Figure 46 illustrates the arrangement of the Tacan ground-beacon antenna. Only the central element, shown as a vertical rod, is excited by the transmitter. This is stationary and has no directivity in the horizontal plane. Actually, it is a vertical stack of discone radiators having a diameter of about 3 in.

FIG. 46. Antenna for ground beacon. Central element fixed in position is energized by transmitter. Parasitic elements are embedded in Fiberglas cylinders rotated 15 rps. A reference-pulse disk is mounted on drive shaft below antenna. (*IT&T.*)

Coarse Bearing. Around the central element is an inner cylinder of insulating material (Fiberglas) with a diameter of about 5 in. This cylinder carries a vertical conductive wire that is completely insulated. The wire acts as a parasitic antenna, or reflector, since it intercepts and reradiates some of the energy radiated by the central element. Its distance from the central element is such that it distorts the normal circular radiation pattern into a cardioid, as shown in Fig. 47*A*. The entire cardioid pattern turns 15 times per sec with the rotation of the inner cylinder.

As a result of this rotation, the signal received at any given direction from the beacon goes through corresponding cyclic variations in strength as a function of time. Figure 47*B* is obtained by replotting the polar-coordinate pattern in rectangular coordinates, with time replacing direction angle as the variable of interest. A cardioid thus transformed into rectangular coordinates is simply a sine wave. For this discussion, the unvarying (d-c) component may be neglected. Since the cardioid is a single-lobed pattern, it takes one full turn of the antenna, or ⅕ sec, for the received signal to go through one complete cycle of variation. In short, the airplane receives a 15-cps modulation that may be demodulated into a sine-wave signal.

A suitable reference signal is required for phase measurement. For this purpose, pulse signals transmitted at some fixed phase and at the correct frequency (15 cps) are utilized. These signals are actually transmitted each instant that the maximum of the rotating cardioid pattern aims due east (for a reason to be explained later). These 15-cps, or coarse-reference signals, are distinguished from the distance-measuring signals by their coding. The distance replies or filler pulses are 12-μsec twin pulses that occur irregularly about 2,700 times per sec; hence the interval between these signals is somewhere in the vicinity of 370 μsec. The 15-cycle reference signal, on the other hand, is a precisely regular group consisting of 12 pulses, each a 12-μsec twin, spaced exactly 30 μsec apart.

This reference signal is illustrated in Fig. 47*B* by a group of extra-close pulses and is for a case of an aircraft due south of the beacon. For an aircraft at a different bearing, the time of arrival of these pulses, relative to the maximum of the 15-cycle pulse envelope, would be different from that shown in the figure.

Reference Signal. An aluminum disk attached to the drive shaft of the rotating cylinder of Fig. 46 has a thin iron slug mounted with its periphery in line with the direction of the maximum of the cardioid-pattern lobe. A stationary pickup coil, mounted close to the periphery of the disk, is due east of the center line of the antenna. Hence, once each turn of the antenna cylinder, at the desired standard instant, a signal is inductively generated in the pickup coil to trigger the distance-measuring transmitter into generating the specially coded coarse-reference signal. On the aircraft, this coded signal is separated from the distance-measuring pulses by a special pulse-group decoder and is then used as the timing reference for the measurement of the phase of the pulse envelope wave.

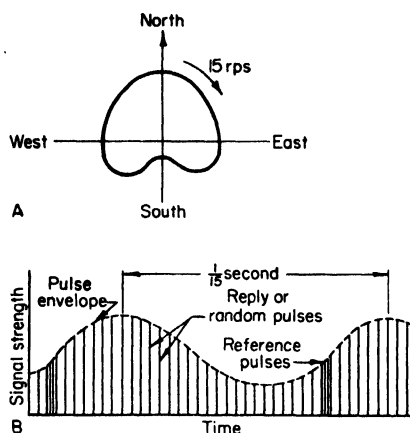



FIG. 47. Circular radiation pattern of central element of antenna is distorted into cardioid pattern at *A* by parasitic element carried by inner cylinder. At *B* the polar-coordinate pattern has been replotted in rectangular coordinates producing a sine wave. The vertical lines represent reply or random pulses, and the group of pulses much closer together are reference signals for phase displacement.

Phase Measurement. In the air-borne receiver, a-f modulation is applied to a calibrated rotary phase shifter, and the output from the shifter is "searched," or moved about in phase, until some preselected index point on the wave is found to coincide in time with the reception of the reference signal. Coincidence is investigated by having a narrow time slot, or gate, formed by the index point on the sine wave, and determining whether the reference signal is received during this time-slot interval. The bearing search may require up to a maximum of 20 sec to complete, depending on the actual bearing at the time. Thereupon the bearing circuits go into track operation. That is, the phase shifter locks to the reference signal and continuously and automatically follows any normal variations in its time of reception with respect to the phase of the bearing-signal envelope. Such normal variations will occur if the bearing is actually changing as a result of the flight path. During the time that the circuits are locked to the reference signal during the tracking operation, the angular position of the rotary phase shifter is a direct measure of the bearing of the airplane with respect to the ground beacon. The angular position of the phase shifter, under proper zero calibration, is used to position the pointer of the bearing meter. This method of operation is stable, accurate, and lends itself to making the bearing data available in the form of a shaft rotation for control of automatic pilots, computers, etc.

The bearing-search process starts automatically each time the air-borne radio set is initially tuned to a new ground-beacon channel or if there is some major interruption in the radio signals. The bearing-measuring circuits also have a memory provision. If, for some reason, bearing signals are lost for a short time up to about 3 sec, the existing bearing indication will be maintained without the search operation being restarted.

All pulses transmitted from the ground beacon, whether intended primarily for distance-measuring reply, filler, station identification, or phase-reference purposes, contribute to forming the a-m pulse envelope or bearing-phase signal. Hence, changing the particular character of the ground-beacon transmissions from time to time in normal system operation does not interrupt or otherwise disturb the bearing-phase signal. Also, radiation of the vital reference signals is arranged to take precedence over any distance-measuring replies, filler pulses, or identity code pulses that might otherwise occur.

The preselected index point actually used for phase measurement is not the maximum of the sine wave, which is a rather broad region, but is the more sharply defined point at which the wave crosses what would be the zero axis in going from its minimum to its maximum value. The index point is ¼ cycle ahead of the maximum, and on the cardioid pattern, it corresponds to a radial index line 90 deg ahead of the maximum. When, in the

course of the rotation of the cardioid pattern, this imaginary index line sweeps past south, the maximum of the cardioid will be momentarily aiming due east. As has been previously stated, the transmission of the 15-cycle reference signal has been standardized to occur at that precise moment.

Bearing Indicators. The bearing meter illustrated in Fig. 44 presents the basic data produced by any omnirange, that is, an absolute bearing or direction angle that is zero-referenced from some fixed geographical direction, generally north (either true or magnetic north, as desired). In line with the type of directional indications that pilots find more directly useful, and to which they are accustomed, the Tacan bearing circuits additionally provide outputs that operate either or both of the standard cockpit indicators, the ID-249 and the ID-250.

On the ID-250 meter, or "radio-magnetic indicator" (Sec. 53), the zero reference is the fore-and-aft axis of the airplane, the top of the fixed meter case (or lubber line) representing the nose of the airplane. To produce this type of indication, the RMI automatically combines Tacan bearing information with magnetic-compass information to actuate the pointer of the instrument. As a result, the angular position of the pointer from the lubber line represents bearing from the aircraft to the beacon measured, not from north, but from the axis of the aircraft. If the meter were to be placed in a horizontal position, the pointer would always physically point toward the beacon, regardless of the heading of the aircraft.

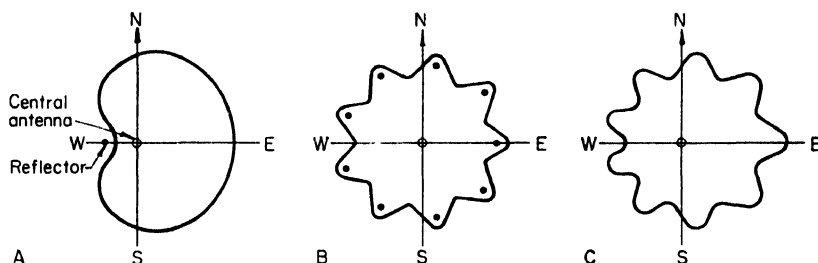



Fig. 48. Antenna pattern (*A*) with single reflector of inner rotary cylinder, (*B*) with nine reflectors of outer cylinder, (*C*) with both inner and outer rotary cylinders.

This relative bearing type of presentation is particularly graphic and convenient for purposes of homing, or keeping the heading of the aircraft aimed directly toward the Tacan omnirange beacon.

On the ID-249, or "course-selector-cross-pointer meter," the center of the fixed meter dial can be considered to represent any selected omnirange radial as the zero reference for the directional indications. This type of presentation is particularly convenient for keeping the aircraft on a given radial track. The pilot sets for a desired radial track by rotating a knob until the selected omnirange bearing figure appears on a numerical indicator. Thereafter, any deviations of the aircraft from the selected radial track are evidenced by sidewise motions of the vertical needle of the cross-pointer meter from the center of the dial. The vertical needle thus directly gives the pilot left-right steering directions for keeping the aircraft on the selected radial. The same principle may be employed for actuating an automatic pilot from Tacan bearing signals.

Fine Bearing. The fine-bearing feature of the Tacan omnirange function produces very greatly improved accuracy over simple cardioid systems. Errors in omnirange bearings arise from two main sources—imperfection of the phase-measuring circuits, which have only some finite limit of accuracy, and, more seriously, radio propagation effects known as site errors. Both errors are significantly reduced by the technique to be described, which requires an increase in antenna dimensions over that for the cardioid or coarse system.

In the case of the Tacan antenna, the necessary increase in size still leaves a compact assembly. Referring to Fig. 46 the inner cylinder has a diameter of about 5 in. The outer cylinder, which is approximately 40 in. in diameter, is constructed of Fiberglas and rotates integrally with the inner cylinder at 15 times per sec. The outer cylinder, however, has 9 wires embedded in it. The wires are spaced uniformly at 40-deg intervals. These wires also are parasitic elements and have a distorting effect on the cardioid pattern produced by the central antenna and inner-cylinder wire. The resulting composite pattern is plotted in polar coordinates in Fig. 48. The over-all cardioid or single-lobed variation is still

present, but superimposed on it are 9 secondary variations or ripples. The maximums of these ripples, or minor lobes, are spaced 40 deg apart.

This envelope signal has a basic $\frac{1}{15}$-sec period, as before, because of the predominant fundamental component due to the rotation of the inner cylinder. The outer cylinder also makes one complete turn in $\frac{1}{15}$ sec, which causes 9 ripples or minor lobes to sweep past the aircraft. Hence the composite envelope wave also contains a ninth-harmonic sine-wave component, having a frequency of $15 \times 9 = 135$ cps. The 135- and 15-cps sine-wave components are separable by filters, and electric phase measurement is performed on each.

To furnish a suitable reference for measuring the phase of the 135-cycle component of the envelope wave, the ground equipment transmits appropriately timed and coded reference

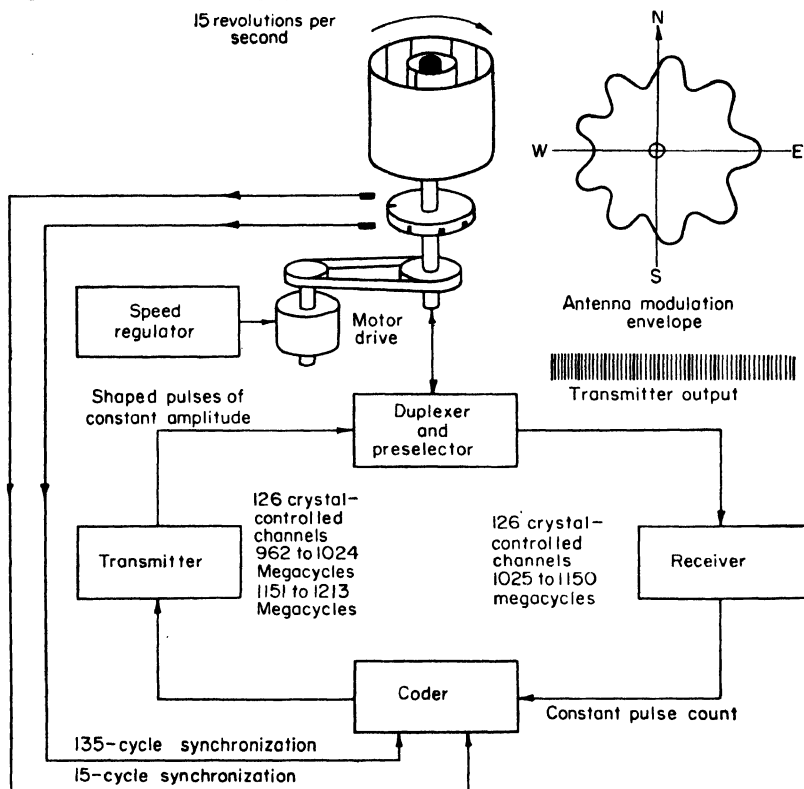



FIG. 49. Block diagram of IT&T Tacan beacon.

signals similar to those for the coarse system. In distinction to both the distance-measuring replies and the coarse-bearing reference signals, the ninth harmonic or fine reference signal is a precisely regular group of 6 pulses, each a 12-μsec twin, spaced exactly 24 μsec apart. In one rotation of the antenna, 8 of these groups are transmitted, separated by 40 deg of rotation. The ninth position is occupied by the 15-cps reference signal.

The important point is that, in the case of the 135-cps signal component, the relation between electric phase and geographic bearing is quite different from that of the 15-cps signal component. Referring to Fig. 48, as an aircraft progresses through a 40-deg arc of bearing, which is the angular width of a minor lobe, the phase of the received 135-cps signal will move through a complete cycle of 360 electrical degrees. Thus, each single degree of bearing change results in a 9-deg change in the measured phase of the 135-cps signal. This ratio of 9 electrical degrees per single space degree gives a pronounced magnifying effect in the process of detecting changes in bearing and forms the basis of the fine-bearing system.

If the bearing indication were given by a pointer that made one dial revolution to indicate 360 deg of bearing, then the circuits should be arranged to produce a 9:1 gearing reduction from the phase measurement to the bearing pointer. One degree of phase shift would then produce $\frac{1}{9}$ deg of pointer movement. If the phase-measuring circuits were accurate to within 1 deg of phase on either side of the correct value, then the bearing-meter indication resulting from measurement on the coarse signal would also be correct to within 1 deg; however, the bearing-meter indication resulting from phase measurement on the fine system would be correct to within $\frac{1}{9}$ deg.

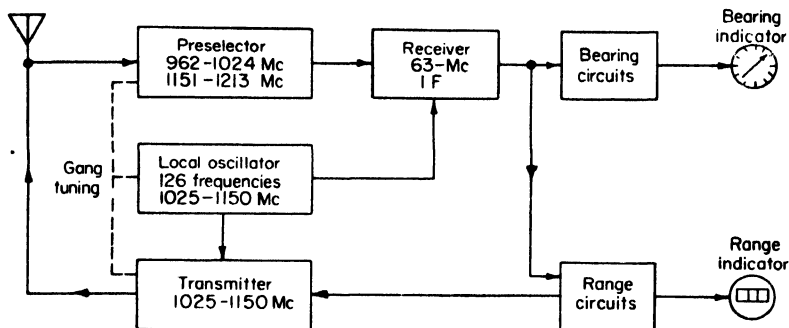



FIG. 50. Air-borne equipment for Tacan.

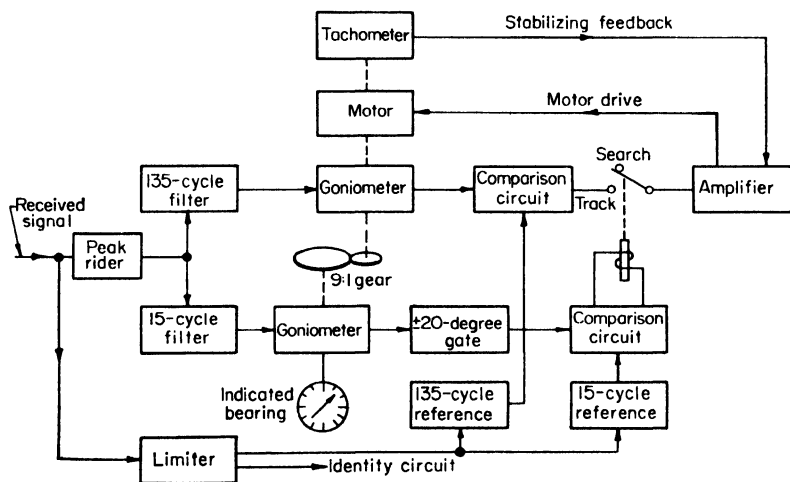



FIG. 51. Air-borne bearing circuits.

To resolve the ambiguity of the 135-cps phase measurement, the 15-cps phase measurement is used. For convenience and simplicity of interpretation, a conventional 360-deg bearing dial with a single pointer is used. Consequently, the 15-cps coarse phase measurement initially moves the pointer to within the correct 40-deg bearing sector, and thereafter ensures that it remains in the proper sector. The exact positioning of the pointer within that sector is under control of the 135-cps fine phase measurement, suitably geared down. The accurate 135-cps bearing information may also be employed for operation of radiomagnetic indicators, left-right meters, automatic pilots, computers, etc.

The improved site freedom of the Tacan omnirange due to the fine system is one of its most useful attributes. With the present equipments, comparative experimental

tests at various locations have shown very substantial reduction in average and extreme bearing errors caused by site effects, as compared with cardioid systems.

The principal components of the Tacan ground equipment are shown in Fig. 49, and follow the principles described above. The air-borne equipment is illustrated in Fig. 50 with details of the bearing and range circuits shown in Figs. 51 and 52, respectively.

74. Coverage and Accuracy of Tacan. The direct-wave component of the space wave is of principal use; however, reflections from the ground or local obstructions are responsible for site-effect errors and for some lobing of the vertical-plane radiation pattern. Both these effects are reduced by the use of antennas which direct more energy above the horizon than below the horizon. The lobing of the vertical-plane radiation pattern is further reduced by the use of vertically polarized energy and by mounting the antenna a large number of wavelengths above the terrain.

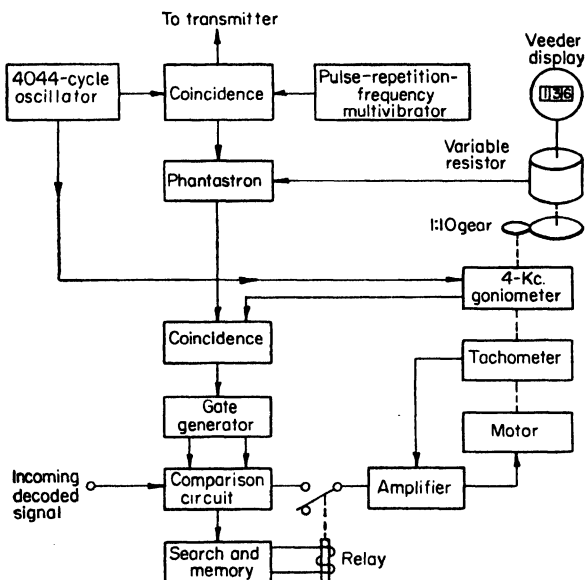



FIG. 52. Tacan range circuit.

At frequencies used for Tacan, coverage and accuracy are not materially affected by factors which usually affect substantially lower frequencies. These factors include electrical ground constants, ionospheric reflections, atmospheric noise, auroral disturbances, precipitation static, etc.

In addition to line-of-sight propagation limitations on coverage, instrumentation circuitry limits the distance-measuring function to a maximum range of 200 nautical miles. The bearing-measuring function has no such limitation; that is, bearing information can be obtained substantially beyond 200 miles when the aircraft is above radio line of sight. With the current standard antenna, bearing information is not available above the ground facility within a total cone of about 90 deg. The angular width of the cone is a function of the antenna design, and new designs are being developed to improve the vertical coverage.

The aggregate bearing error of Tacan equipment currently in service is typically ± 1 to ± 2 deg if the equipment is adjusted so that the mean error is zero. Field tests of modified equipments (both ground and aircraft) have demonstrated the feasibility of errors (excluding attitude-effect errors) within ± 0.5 to ± 1 deg. Future production equipments should be capable of at least this accuracy.

The distance-measuring error of the Tacan system is nominally specified as ±0.2 nautical mile, regardless of range. Modern Tacan equipment operates well within this specification.

Unlike the bearing-measuring accuracy, the distance-measuring accuracy is not appreciably affected by siting or by the reception of a horizontally polarized component.

PART 3

THE EQUIPMENT OF THE APPROACH AND LANDING ZONE

The equipments to be described in this section are those utilized in navigating within the approach and landing zone, *i.e.*, the terminal area. This area has been designated as an airspace of defined dimensions which accommodates aircraft arriving at, departing from, and operating in, the vicinity of airports. It was defined by Special Committee 31 of the RTCA as consisting of the final-approach and live-runway area, the initial approach area, the holding area, and the departure area. The zone[68] is generally considered to consist of a radius of about 30 miles and extending upward from sea level to an altitude of 7,000 ft. These dimensions may vary, and there are reasons to believe[69] that with higher-speed aircraft the altitude may have to be extended. In a large metropolis, several airports forming a single airport system may be located within a common zone.

As in the short-distance zone, the major problem in the terminal area is air-traffic control, but it is more severe in nature since traffic densities are higher. Hence, it is important that the navigational system of the terminal area be capable of producing a controlled safe separation between aircraft that is not wasteful of space and therefore time. With navigational systems utilized in the prewar period, it was necessary to maintain spacings equivalent to 8 min at touch-down. During good weather, however, landings without instruments were made at intervals of less than 30 sec. Attaining safe separations equivalent to this short time interval is the goal of modern electronic navigation systems intended for use in the approach and landing zone. With present operational terminal-area equipment, spacings equivalent to about 2 min have actually been achieved.[70]

Another important distinction of the terminal area is that navigational information must apply in three dimensions. In the two en-route zones, the vertical position of an aircraft is held by reference to a barometric altimeter and navigation information is provided on a two-dimensional basis. In the terminal area, particularly during approach, the altitude must be very carefully controlled if safety is to result.

It is the present philosophy, in air-traffic control for the terminal area, that the burden of maintaining safety rests with personnel stationed on the ground. This provision necessarily requires a large volume of reliable communications and forms a vital part of present traffic-control systems. The communication equipment is covered elsewhere in this handbook, the following section being limited to the electronic aids to aviation.

75. Airport Surveillance Radar. Although radar was first developed for early-warning purposes in the detection of enemy aircraft, it was soon applied to a variety of navigational systems for use with friendly aircraft. The requirements for a surveillance radar were discussed by the Communications Technical Division of the PICAO in 1947,[71] but a firm recommendation for such equipment did not appear until the fourth meeting of the Communications Division held in 1951.[72] When used for air navigation and traffic control, it is essential to distinguish between pure radar systems and those involving interrogators and transponders. Pure radar utilizes only the natural reflection from the target and is termed *primary radar*. The use of a transponder at the target (such as an aircraft) greatly increases the reliability of the "echo," and such a system is called a *secondary* or *cooperative radar*.

Special Committee 31 of RTCA recommended[73] a secondary-radar system for the control of traffic in the ultimate equipment. For the interim period, however, it suggested the implementation of an extensive net of surveillance radar for airport-terminal-area use. It also favored, during this interval, the installation of air-borne beacons that would in effect convert the primary to secondary radars. Such radar installations were made beginning in about 1949, and in the fall of 1950, a meeting was held by the Operational Policy Group of the Air Traffic Control and Navigation Panel of the Air Coordinating Committee in the United States to determine how these equipments could be employed. The recommendation[74] of this group was as follows: "Radar and the radar controller should become the primary means of controlling air traffic at congested terminals. Of all of the aids now becoming available, airport surveillance radar is capable of making the most significant contribution towards increased safety and efficiency of operation. There are a number of sound reasons why the terminal-area traffic-control system should be built around the surveillance radar."

The ACC group also recommended air-borne beacons to cooperate with the ground radar.

Although the operating principles are the same, there are many differences between a ground surveillance radar and air-borne electronic pilotage equipment. The outstanding difference pertains to the source of reflection. With the air-borne equipment, reflection is obtained from an entire area having a diameter of perhaps 100 miles, whereas with the ground equipment the total reflecting surface may be a single aircraft located many miles from the terminal radar. Also, the air-borne equipment requires a maximum range of perhaps 10 miles (the altitude of the aircraft), whereas the ground equipment requires several times this value. The definition must also be improved in ground radars since it is called upon to distinguish safely between two or more closely spaced targets.

The effect of weather on ground radars is more severe than encountered in electronic pilotage equipment because of the normal incidence of propagation with respect to the atmosphere. In air-borne radar, radio waves travel at nearly vertical incidence and are seldom attenuated to a point where they do not yield information on the configuration of the terrain. When used at horizontal angles, the air-borne equipment constitutes a weather radar providing valuable storm information to flight crews. On the other hand, rays from the ground equipment must penetrate this same weather at nearly horizontal angles and be capable of distinguishing a small target.

In the application to traffic control, each target appearing on the screen of a ground radar requires the attention of a controller, and in areas of high traffic density, this requires that the total problem be divided among several persons. Accordingly, it is often necessary to provide a multiplicity of displays.

In general, surveillance radar must utilize high directivity to yield accurate directional information, the p.r.f. must be high to afford indicator persistence, the pulse length must be short for good range resolution, and the peak power must be high to give adequate range. Because of the usual requirement for small antenna structures, the trend is toward the higher frequencies, even though these may exhibit some problems in propagation and high attenuation in certain weather. Surveillance radars operate largely at frequencies in the vicinity of 1,500, 3,000, and 10,000 Mc. A typical example of a primary ground surveillance radar is the search system of GCA equipment, which is described in Sec. 87.

MTI. Primary ground radar, unlike air-borne radar, is limited in close-range operation by permanent echoes caused by terrain features, clouds, or surfaces near the antenna, *i.e., ground clutter.* Since it is desirable to locate a surveillance radar near the airport, it is often not possible to choose sites for best freedom from permanent echoes. A noteworthy contribution to this problem is the moving-target indicator (MTI). The principle employs the Doppler effect, which has been discussed in Sec. 47. Briefly, a simple form of MTI operates in the following manner. The ground transmitter utilizes a master oscillator supplying c.w. and a power amplifier to which the pulsing is applied. The pulsed echoes are then beat with the output of the local mas-

ter oscillator (sometimes called the *coherent oscillator* or *coho*), and the existence of a Doppler frequency indicates that the reflecting object is moving with respect to the radar station. Since permanent echoes result from stationary targets, the beat frequency in these cases will be zero. Hence, by using a high p.r.f. and an appropriate filter, the desired reflection from a moving aircraft can be separated from the undesired reflections or permanent echoes. It should be noted that, if the aircraft is moving at right angles with respect to the line joining the aircraft and the radio station, the beat frequency will be zero and the aircraft will not be seen. Hence, MTI will work only when the rate of closure is not zero. This is characteristic of all MTI circuits.

Displays. Additional displays are often necessary to best utilize the capabilities of a ground surveillance radar. The *plan-position indicator* (PPI) is a popular type of display used in all surveillance radars; however, it has certain weaknesses that are remedied by the use of additional displays. The PPI employs a radial time base with zero time at the center of the cathode-ray display. It constitutes a polar plot with angles measured at the center of the tube representing bearings to various targets that appear on its face. Essentially, therefore, it constitutes a maplike representation. If the scale is not sufficiently large, it may lack the accuracy required by a controller to

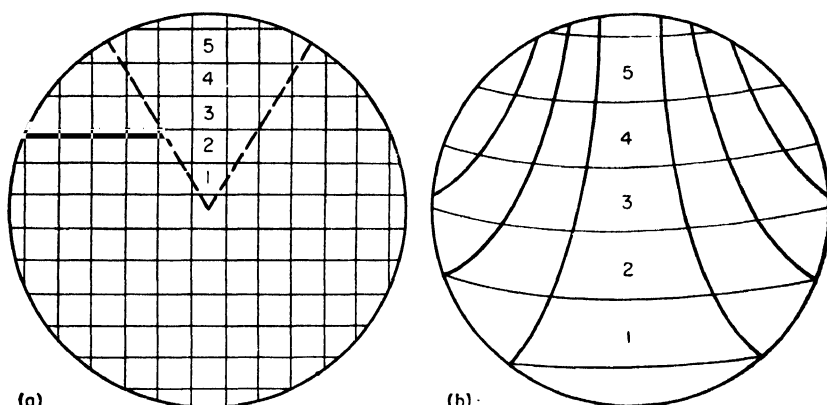



FIG. 53. Sector shown at (*a*) for standard PPI display appears as in (*b*) for a *B*-type display.

notice the existence of a dangerous condition or to determine, with sufficient accuracy, what vectors are to be issued to aircraft. Accordingly, two other displays are often used, normally called the *off-center* PPI and the *B display.*

The principles of the off-center plan-position indicator can be visualized by imagining a PPI having a diameter of, say, 12 ft. However, the radar operator has at his disposal an indicator of only 15 or 20 in. in diameter. A control is provided which permits the operator to place the center of the small indicator anywhere he wishes on the face of the large imaginary PPI. The operator, therefore, has available any portion (but not all) of the comprehensive display, with a scale that may be more than ten times the scale of the PPI normally at his disposal.

The off-center plan-position display has its greatest usefulness in increasing the definition of areas located some distance from the radar. However, increased definition is more frequently required near the radar, and for this purpose the *B*-type display offers many advantages. Whereas PPI is a polar-type display, the *B*-type is quasi-rectangular, and the two are compared in Fig. 53. At *a* is shown the face of a cathode-ray tube that might be used for a PPI. The face of the tube is layed out in square areas, and a sector having an angular width of about 60 deg is laid off in dashed lines. The appearance of a *B* display covering approximately the same sector as shown in Fig. 53*a* is shown at 53*b*. The numbered areas in the two displays correspond. The improvement in resolution obtained by the *B* display, particularly in the area that

would correspond to the center of the PPI, is evident. However, directions on the *B* display do not represent bearings as accurately as do those on the PPI.

Height Finding. In some traffic-control applications, it has been found desirable to associate height finding with the surveillance radar. Height finding is employed largely for the purpose of distinguishing between two aircraft in the same area. By communication, the altitude of the aircraft is determined and a height-finding determination is made. This process then serves to identify the target at the altitude corresponding to the communicated information. Height finding is seldom of sufficient importance in surveillance radar to demand elaboration of the basic equipment but is normally accomplished by use of a separate small radar.

The height-finding radar makes use of an antenna that is a paraboloid of elliptical aperture. The antenna is so mounted that its long dimension is vertical. Therefore, it has a narrow pattern in the vertical plane and a broader pattern in the horizontal plane. The antenna oscillates in the vertical plane so that the position of the beam varies from an angle of 2 deg below the horizon to 23 deg above the horizon. This nodding cycle occurs at a rate of about once per second.

Height finders have been constructed to operate at frequencies of 3,000 and 10,000 Mc. The higher frequency is preferred because it is possible to obtain a narrower beam with an antenna of reasonable size.

76. Fixed-beam Low-approach Systems. An instrument low-approach system (ILS) comprises a means for furnishing guidance in the vertical and horizontal planes to aircraft during descent from an initial approach altitude to a point near the ground. Completion of a landing requires guidance to touch-down by means other than the approach system, and when this is provided the combination may be properly termed a landing system.[75] The use of shaped radiation patterns to provide this guidance were first proposed by H. Diamond and F. W. Dunmore[76] of the U.S. Bureau of Standards in about 1928. Early in the Second World War, the United States Air Corps took over the system that the CAA had been developing,[77,78] and produced equipment that was installed throughout the world. The resultant system, developed by the Federal Telephone and Radio Corporation, bore the nomenclature SCS-51.[79] After the war, ICAO adopted the SCS-51 as its standard.[80] In subsequent meetings of the ICAO modifications were made in the original specification but these were largely concerned with more accurate definitions and methods for adjustment of the equipment. Essentially, the basic system originally adopted by that organization has continued to be maintained as a world standard.

A fixed-beam low-approach system consists of three basic components: the *localizer*, the *glide slope*, and *marker beacons*. The approach procedure requires that the aircraft be guided both laterally and vertically, and that some indication of distance be given. Accordingly, the localizer provides lateral guidance, the glide slope provides vertical guidance, and marker beacons are used to indicate the distance to the point of touch-down.

In a standard fixed-beam low-approach system, the localizer transmitter operates between 108 and 112 Mc and is installed at the far end of the runway (opposite from the approach end). It is located so that the center of its antenna array coincides with the center line of the runway. It transmits shaped patterns, to be described later, which essentially define the extended center line of the runway and thereby provide the required lateral guidance.

The glide-slope equipment operates in the 332.6- to 335-Mc band and is located at a distance of approximately 750 ft from the approaching end of the runway and approximately 450 ft to either side of the center line of the runway. This equipment generates a signal which defines a plane inclined with respect to the horizontal and thereby provides guidance in the vertical plane.

The localizer and glide-slope signals are received in the aircraft by means of separate antennas and receivers. The navigational information is displayed on a cross-pointer instrument which incorporates two indicating needles, one pivoted from the top and the other from the side. The position of the vertical needle (which deflects right or

left) indicates the position of the aircraft with respect to the localizer course. Similarly, the position of the horizontal needle (which moves up or down) indicates the position of the aircraft with respect to the glide slope. The aircraft is on the proper approach course when the two needles cross at the center of the instrument.

To indicate progress along the approach path, use is made of three marker beacons normally referred to as the boundary marker, the middle marker, and the outer marker. These markers operate at 75 Mc and are identical with those described previously in Sec. 22. The power of the beacons employed with approach systems is normally of the order of 3 watts. The three markers are distinguished from one another by the utilization of a different modulating frequency. Normally, ILS installations

FIG. 54. Localizer system employing mechanical modulation.

of the CAA in the United States utilize only the outer and middle markers. The boundary marker, if used, is located at a distance of approximately 250 ft from the approach end of the runway, and is modulated at 300 cps. The middle marker, normally located approximately 3,500 ft from the approach end of the runway, is modulated at 1,300 cps. Finally, the outer marker, located at a distance of approximately 4½ miles from the touch-down point, is modulated at 400 cps.

Details of the marker beacon have been previously described; hence the following sections will be limited to a discussion of the localizer and glide-slope components.

77. The Localizer Equipment. A typical standard localizer system is shown diagrammatically in Fig. 54. Some of the systems have used V-type horizontal antennas instead of the Alford loops shown. Arrays have been constructed employing from five to eight loops, but basically the principles remain the same.

The unmodulated transmitter output is connected first to an r-f bridge termed the *cross-modulation bridge.*' The bridge arms are $\lambda/4$ or less in length, and because of the

cross connection at H, the r-f energy from both arms of the bridge arrives at G out of phase and therefore cancels. The adjustment at K is made to balance the bridge, and when so balanced, the bridge serves to divide the total r-f energy into two equal portions and prevent interaction between these portions.

Mechanical Modulator. Modulation has been accomplished by both electronic and mechanical means. Figure 54 employs the mechanical method, which has exhibited a very high order of stability. The mechanical modulator consists simply of $\lambda/4$ transmission-line sections coupled to the lines that lead from the cross-modulation bridge to the antenna bridge. The high-impedance ends of the coupled sections terminate in capacitors. In the center of each capacitor are heavy capacitor plates rotated by a motor. When the coupled sections are tuned, they reflect a virtual short

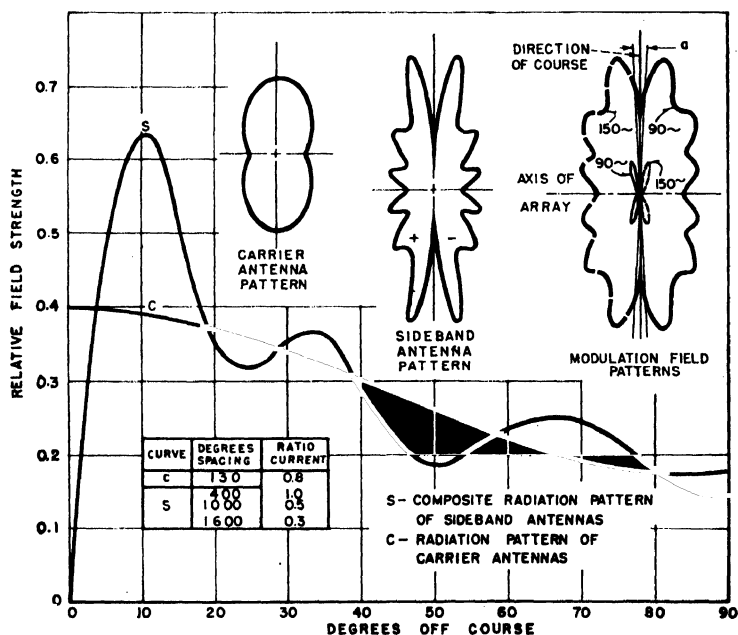



FIG. 55. Directive field pattern for equisignal localizer (ILS).

circuit in the transmission line. As the rotary plates are engaged, the coupled section is detuned and the short circuit is thus removed. By properly shaping the rotating capacitor plates, it becomes possible to effect a sinusoidal modulation. The mechanical modulator does not add energy and thereby constitutes a downward modulation. With three and five plates, respectively, being rotated at 1,800 rpm, the required 90- and 150-cps modulation components are generated.

Because of the modulation, the input to C on the antenna bridge consists of carrier frequency and side bands of 150 cps. Similarly, at corner B, the input consists of carrier frequency plus side bands of 90 cps. Because of the transposition at D, the carrier frequency will cancel at E, leaving only side bands of 90 and 150 cps. The total carrier energy, plus half the side-band energy, is present at the opposite corner of the bridge and furnishes the excitation for the central antenna of the array. By proper adjustment of cable lengths, the side-band components fed to the side-band antennas are 90 deg out of phase with the corresponding components in the central antenna. A phase reversal is employed at F in the common transmission line joining each set of side-band antennas. Therefore, the side bands on the right-hand side of the array

will be 180 deg out of phase with the corresponding side bands on the left-hand side of the array.

The phasing and distribution discussed above produce directive patterns typified in Fig. 55. Since the side-band energy is symmetrically distributed but out of phase on both sides of the array, these antennas contribute a null along the perpendicular bisector of the array. Hence, along this line, the only energy received by an air-borne

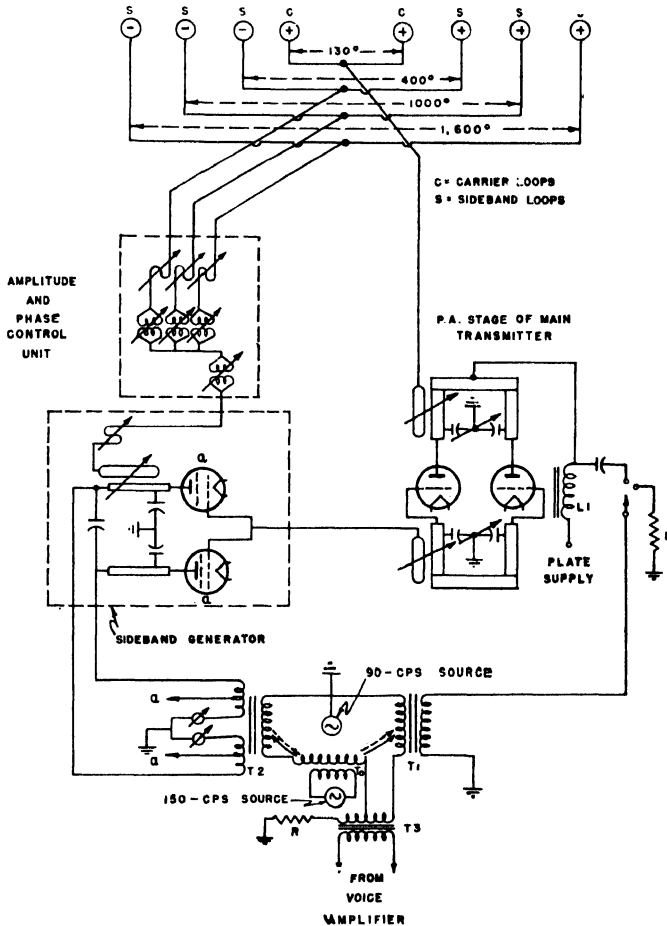



FIG. 56. Localizer system employing electronic modulation.

equipment is that radiated by the central radiator. When the carrier and side-band components of the signal are combined, the resultant modulation field pattern yields virtually 90 cps alone on one side of the runway and 150 cps alone on the other side. The course is defined by the intersection of these patterns, at which point equal magnitudes of 90 and 150 cps are obtained. For this reason, this type of equipment has been termed the *equisignal localizer.*

Electronic Modulator. A typical electronically modulated localizer system is shown in Fig. 56. The principles of operation are generally similar to those described above except that an electronic side-band generator is utilized instead of the mechanical modulator. The side-band generator consists of a balanced modulator in which the

carrier component is canceled in the common plate circuit, leaving only the side-band components. These components are connected to the side-band array as described for the mechanical system. The modulation tones are also applied to the output stage of the main transmitter, which comprises the excitation for the central array. In Fig. 56, two antennas in parallel are utilized for this purpose. Since the equality of 90 and 150 cps determines the course position, it is vital that these two tones be carefully generated and kept balanced in the system.

Two important characteristics of a localizer are the course width and clearance. The course width determines the deflection obtained for a given deviation from the course line. ICAO standards[81] have specified that the localizer must cause a full-scale deflection of the needle when the aircraft has deviated 2.5 deg from the center line. This deflection is a function of the difference between the magnitudes of the 90- and 150-cps components at the output of the radio receiver. Hence, the indicated course width depends upon proper calibration of both the air-borne and ground equipments.

The ratio of amplitude of the two modulation frequencies in any direction, expressed in decibels, is defined as the clearance. Hence the clearance is zero on the course line. For angles up to 2.5 deg on either side of the center line, the meter deflection is very nearly proportional to angular displacement; but for all other angles the deflection should never be less than full scale and should always be in the correct sense. In normal operation, as soon as the needle leaves full-scale deflection, the pilot regards the indication as signifying that the aircraft is within 2.5 deg of the center line and therefore begins to correct the heading accordingly. Hence, should the needle leave full-scale deflection at other regions besides the course region, some confusion may result. Any region beyond 2.5 deg from course in which the needle is not full scale is therefore called a "low clearance region."

One of the most important characteristics of localizer courses is the amount and character of their bends. An excessive amount of certain bends will render a course totally unusable since it becomes impossible to follow it accurately. The bends are caused by reflecting obstacles in the vicinity of the transmitting equipment, and several techniques have been developed for minimizing them. If the obstacles are located generally behind the transmitting equipment, one technique is the use of a reflecting screen which reduces the energy directed at the offending obstacles. In general, however, this technique produces an unflyable back course. A preferred technique is the use of higher directivity and increased sharpness at the intersection of the major lobes. It can be shown that the bending that occurs near the center line of a localizer for a given condition of reflected signal is directly a function of the gain setting of the receiver. However, to retain the same apparent course width, the difference in the side-band field at a given bearing must be increased. This corresponds to saying that the angle of intersection between the two lobes must be decreased so that a given deviation from the on-course position provides a greater difference in the modulation tones. Normally, the generation of a sharper intersection is coincident with more directivity in the lobe structure. Accordingly, less energy is directed in regions remote from the course where reflecting obstacles are more likely to be. This additional sharpness can be obtained by the use of more antenna elements occupying a wider aperture. For this reason, a special localizer equipment designed to produce acceptable courses in poor sites has been termed the "wide-aperture localizer."

78. The Wide-aperture Localizer. Recognizing the need for improved localizer performance, in about 1950 the CAA and the United States Air Force gave consideration[82] to the development of a wide-aperture localizer system, and equipment of this type emerged from production early in 1954. This equipment carries the military nomenclature AN/MRN-7.

The wide-aperture localizer had, as a design goal, a 10:1 improvement in freedom from site effects. The antenna developed for this purpose consists of a screen about 85 ft long and 7 ft high, which mounts 12 horizontal dipoles $\lambda/4$ from its surface. The spacing between centers of all dipoles (except the end elements) is 268 electrical degrees at the central frequency of the localizer band. The last elements are

spaced 308 deg from the centers of their nearest neighbors. All elements, except the two on the extreme ends, are fed carrier energy; hence, an extremely directive carrier pattern results. Side-band energy is fed to all elements of the array with the usual phase reversal between pairs on each side of center. The phase reversal is accomplished by connecting the pairs through transmission-line bridges similar to those previously described. The bridge further serves as a hybrid transformer to isolate the antenna pairs. The r-f inputs are then applied to the two remaining corners of the bridge and permit the simultaneous excitation of the two antennas of the pair with

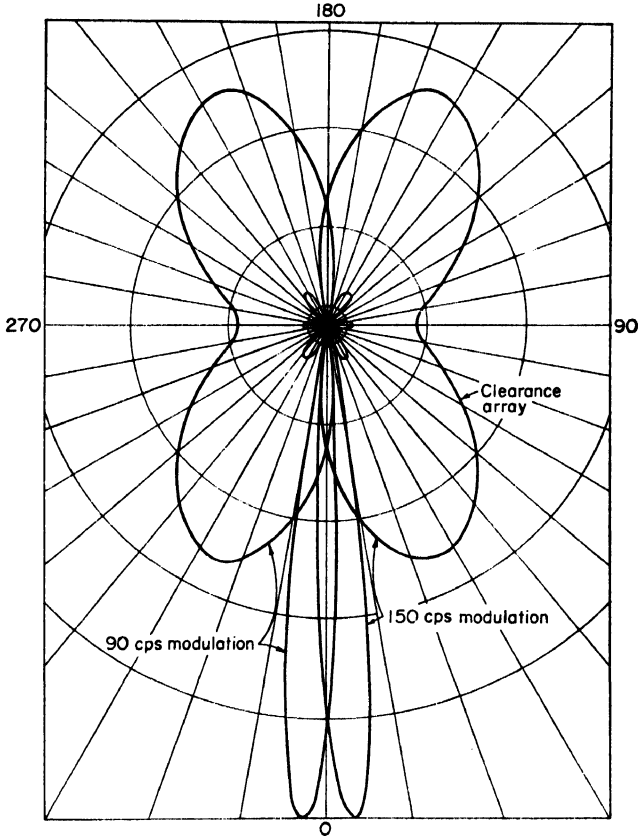



FIG. 57. Field pattern of a wide-aperture-localizer system.

carrier signal in phase and side-band signal out of phase. Existing transmitters and modulating equipments can be used with this array. One terminal of each bridge (except that for the end element) is connected to terminal A on the bridge shown in Fig. 54. The other terminals of all the bridges are connected to terminal E of the antenna bridge shown in Fig. 54 so that all the antennas are supplied with side-band energy. The bridges are connected to coaxial transmission-line transformers, by which it is possible to match impedances and control the amount of current to the pairs. The resulting side-band pattern of the array is shown in Fig. 57. It is evident by comparison with the normal localizer patterns shown in Fig. 55 that a substantial increase in directivity and pattern sharpness has been attained. Hence, reflecting obstacles to either side of course now receive less side-band energy compared with the

signal in the vicinity of course, and less bends result. To meet the 2.5-deg course-width requirement using a standard localizer receiver, the normal percentage of modulation is adjusted downward so that a standard receiver can fly a normal localizer or the wide-aperture localizer interchangeably.

With the narrow field pattern shown in Fig. 57 it would not be possible to obtain guidance except for a very narrow sector having a total width of approximately 12 deg. To overcome this difficulty, use is made of an auxiliary array employing only three elements. This assembly is called the *clearance array* and its pattern is also shown in Fig. 57. It is located about 50 ft back of the main array and is operated from a separate transmitter having a carrier frequency 10 kc removed from the frequency of the transmitter that excites the main array. The band width of the localizer receiver is such that it readily receives both the clearance- and precision-array signals without retuning. As the aircraft passes from the regions where it receives only clearance signals to the approach position, the signal from the precision array is so much greater that the clearance array no longer contributes to the indication. This phenomenon has been termed "capture" effect and ensures that in the vicinity of course only the precision array is effective.

79. Localizer Apparatus. The ICAO specifies that localizer service over a sector of 20 deg from course shall be provided to a distance of at least 25 nautical miles from the approach end of the runway at an altitude of 2,000 ft. It is also required that, in all other sectors, guidance be provided to a distance of at least 17 nautical miles at an altitude of 2,000 ft. These requirements can be met readily with transmitter powers of from 100 to 200 watts. Actually, the requirements can be met with a radiated power of approximately 25 watts, but if a mechanical modulator is used it must be remembered that its downward modulation does not permit radiation of the total power.

Transmitters have been of conventional design, utilizing crystal control at approximately one-eighteenth of the output frequency, followed by conventional frequency multipliers. Older designs employed balanced circuits because it was believed that balanced excitation of the antenna was necessary to avoid polarization effects. Modern antenna arrays and transmitters, however, have been developed for operation with coaxial lines. Original transmitters, used with mechanical modulators, employed five tubes; but seven tubes (exclusive of audio oscillators and tone generators) are necessary when electronic modulation is employed. Modern transmitters employing coaxial design use only four tubes. Because of the very high gain of wide-aperture arrays, a transmitter radiating 15 watts of power is sufficient for use with this type of antenna system.

ICAO specifies that the localizer shall provide for the simultaneous transmission of an identification signal. It has been the practice in the United States to utilize the identification circuit as an additional ground-to-air communications channel. Tone identification is provided at approximately 1,020 cps. This tone is keyed with a Morse signal by an automatic keyer.

Transmitters employing mechanical modulation have sometimes made use of a separate channel for tone-identification keying. When this technique is employed, an additional r-f bridge is inserted which connects between the antenna bridge of Fig. 57 and the central antenna. The third terminal of the bridge is connected to a modulated r-f amplifier, while the fourth terminal connects to a matching load. It is practical, however, to voice-modulate the main transmitter (which is also mechanically modulated with 90- and 150-cps tones) without employing a separate r-f amplifier. In this case, hum is evident in the voice transmission but the intelligibility has been considered adequate.

During the war years, a radio receiver known as the BC-338 was employed for the reception of localizer signals and provided for six channels. During the postwar period, it was realized that the receiving systems of the localizer and VOR could be conveniently and advantageously combined. Accordingly, receivers produced in the United States for the reception of localizer signals are common with the VOR receiver. Refer to Sec. 50 for remarks concerning this equipment. When the VOR receiver

is tuned to one of the channels designated for localizer operation, the phase-comparison circuit normally used for VOR is automatically replaced with the localizer amplitude-comparison circuit in which the output of the receiver is separated into individual 90- and 150-cps components by filters. The output of each filter then goes to a separate bridge rectifier consisting of a series of copper-oxide elements. These elements are temperature-sensitive, and to prevent the course width from changing as the temperature rises, use is made of a disk-type thermister of high negative temperature coefficient. The outputs of the two rectifiers are then connected in opposition to the localizer instrument. A typical amplitude-comparison circuit is shown in Fig. 58.

As the schematic of Fig. 58 indicates, a voltage is also developed which is proportional to the total of the rectified 90- and 150-cps signals. This voltage is used to actuate a separate coil in the indicating instrument that causes a small tab to disappear from the face of the meter. This tab, commonly called a flag, bears the printed notation "off." Therefore, if insufficient signals are present, the pilot is warned to disregard the indication of the meter. This alarm is vital because, while equal amounts of 90- and 150-cps signals cause the indicator to be centered, the same indication may be obtained if the equipment fails and neither signal is present.

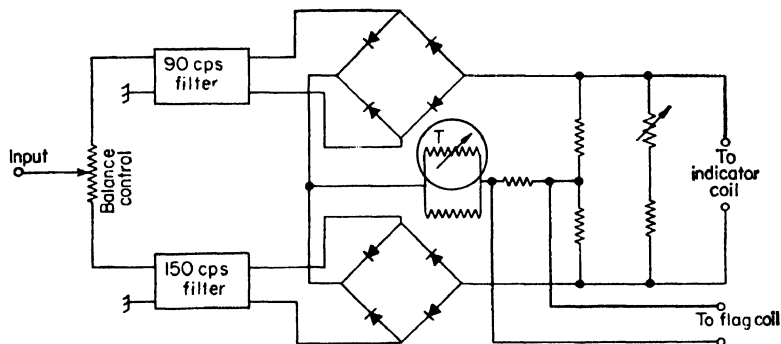



Fig. 58. Circuit for deriving localizer or glide-slope indication.

Errors. There are two main categories of errors attributable to the air-borne receiving equipment, and these consist of the course position and the course width. Initial calibration adjustments are available for both these quantities. With aging, however, some differences in attenuation may occur in the 90- and 150-cps filters, or in the associated copper-oxide rectifiers. The stability of the course width, once adjusted, is mainly dependent upon proper operation of the AGC. To maintain high standards of safety during the critical landing phase, course position and course width are checked at frequent intervals as a normal preventive-maintenance procedure.

Course-position errors are also possible because of distortion of the 90- and 150-cps filters. For this reason, the modulating equipment on the ground and the audio amplifiers in the air-borne equipment are designed for a high order of fidelity.

80. The Glide-slope Equipment. The *equipotential* or *constant-intensity glide slope* attributed to F. H. Dunmore of the U.S. Bureau of Standards plays no part in standard fixed-beam low-approach systems adapted by the ICAO; however, slopes employing this principle were important in all low-approach developments in the years from 1927 to 1941. This type of glide slope, despite several ingenious later modifications, had one or more of the following defects to an undesirable degree.

1. Lack of range or inability to produce a straight-line path more than a few miles in length
2. Lack of uniformity in the glide-slope angle from one airport to another and from one aircraft to another, due to differences in transmitted energies or receiver sensitivities

d. Change of the glide-slope angle depending upon the altitude of the aircraft with respect to the transmitting antenna

81. The Equisignal Glide Slope. These difficulties of the equipotential glide slope are overcome by the equisignal-type glide slope which was first tried in 1938 by D. M. Stewart and J. C. Hromada of CAA. The system, accepted as standard by ICAO, was developed by the Federal Telecommunications Laboratories for the United States Air Force in about 1941.[83] The general principle of operation is similar to the localizer in that overlapping field patterns, carrying different types of modulation, are employed. The position of the slope is therefore determined by the equality of audio tones.

A basic glide-slope system is shown in Fig. 59. The crystal-controlled transmitter operates in the band from 329.3 to 335.0 Mc. The output of the transmitter is handled in a manner completely identical to that of the localizer previously described, except that the antenna bridge is no longer used. Energy modulated at 150 cps is connected to an upper antenna and energy modulated at 90 cps is connected to the lower antenna. Attention is called to the amplitude control whereby the relative amount of energy of the two antennas may be adjusted. For a 2-deg glide angle, the upper antenna is placed at 28.5 ft above ground and the lower antenna at 6 ft above ground. By decreasing these heights while keeping the height ratios constant, the glide angle may be increased to 5 deg in steps of ¼ deg to 4 deg and steps of ½ deg up to 5 deg. With this arrangement, the relationship of the multilobe patterns produced by the two antennas is as shown in Fig. 60. The pattern radiated by the upper antenna has several narrow lobes within the space occupied by the first wide lobe of the lower antenna. Inasmuch as the slope position is determined by the intersection of the upper and lower antenna patterns, only the lowest of the narrow lobes can be permitted to intersect the wider lobe; otherwise multiple courses would be produced. Avoidance of these other intersections is achieved by properly limiting the amplitude of the field from the upper antenna compared with that of the lower antenna. This is accomplished by the amplitude control referred to previously. Too small a signal from the upper antenna, however, results in a somewhat indefinite intersection, as shown in Fig. 60. Even with the intermediate value shown, a false course is produced at the point a', namely, 20 deg above the horizontal. At this high angle, however, there is little chance of confusion since the glide angle is some six to ten times the operationally desirable angle and would be difficult if not impossible to follow.

FIG. 59. Equisignal glide-slope system.

The foregoing analysis has treated the glide slope as though it extended from the region on the ground directly below the two antennas, the equisignal area being in the shape of an inverted cone. Actually, however, the glide-slope equipment is located at a distance of 400 ± 150 ft from the center line of the runway to avoid obstruction to the approach. With this arrangement, the equisignal path along the runway takes the form of a hyperbola, that is, the cross section produced when the cone is cut by a vertical plane other than that passing through its apex. Since the object of the installation is to provide a path which is substantially straight throughout its length, it becomes necessary to counteract the above-mentioned effect. The slope can be changed by adjusting the ratio of the amplitudes of the

field patterns produced by the two antennas. It should be noted that, when viewed from the glide-slope station, the aircraft in its descent subtends a considerable horizontal angle. By employing slightly different directional characteristics in the horizontal plane of the two transmitting antennas, the amplitude ratio between the upper and lower antenna patterns is made to change with horizontal angle from the station by the amount required to produce a straight-line glide slope. The desired effect is obtained by orienting the upper antenna so that its direction of maximum radiation is displaced 12 deg from that of the lower antenna which is parallel with the runway.

The lower antenna consists essentially of half an Alford loop used in conjunction with a reflecting screen, the combination being the equivalent of a full loop except that the transmitted energy decreases rapidly behind the screen. The pattern in front of the screen, however, is nearly circular. The upper antenna is in the form of two V-shaped horizontal dipoles mounted directly above each other at a spacing of λ/2. Both the upper and lower antennas are held by a 3-in. aluminum mast which is mounted on, and guyed to, a small enclosure housing the transmitting equipment.

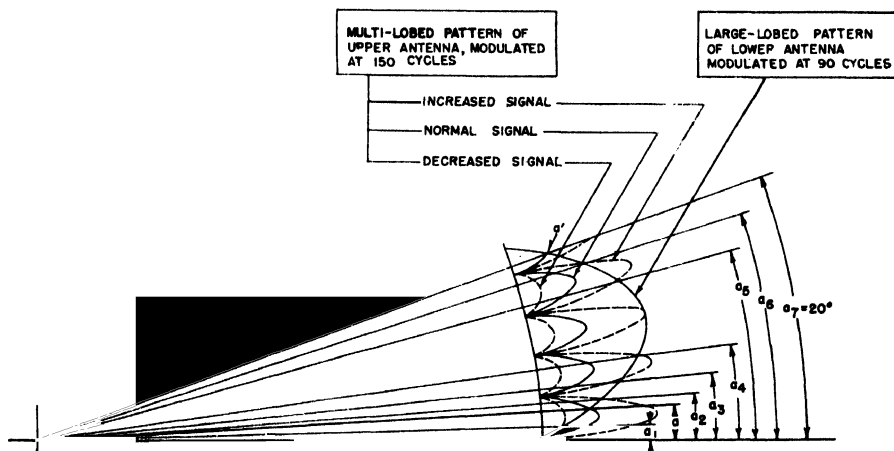



FIG. 60. Production of an equisignal glide path by intersecting field patterns radiated from two antennas at specified different heights above ground (ILS).

82. The Null-type Glide Slope. At high approach angles, it is necessary for the lower antenna of the equisignal glide slope to be located in close proximity to the ground. Since the patterns are formed by ground reflection, it is easy to understand why heavy snowfall (which reduces the effective height of the antenna above ground) will cause the path angle to change. Recognizing this difficulty, the CAA undertook the development[84] of a null glide-slope system in about 1947. The first equipment embodying this principle was produced under government contract by the Federal Telephone and Radio Corporation as the type TUS glide-slope equipment.

The principle of this equipment is shown in Fig. 61. Similarities can be noted with the equisignal-type system of Fig. 59. The null-type system, however, incorporates an antenna bridge and, in that respect, becomes entirely similar to a typical localizer system. The corner of the antenna bridge which supplies side bands only is connected to the upper antenna while the opposite corner, supplying carrier and side bands, is connected to the lower antenna. There is, in addition, a third antenna connected to the side-band corner, referred to as a "modifier antenna." Its purpose will be discussed later.

The vertical patterns produced by the three antennas are shown in Fig. 62. These patterns indicate that the upper antenna produces a null at the desired glide-slope angle. The patterns have been calculated for an upper antenna height of 4,000 electrical degrees, which corresponds, at 330 Mc, to approximately 33 ft. The two lobes on either side of the glide-slope null contain only side bands, but their phases are in opposition. This phase relation obtains naturally from the ground reflections and can be compared with the phase relation of the right and left side-band antennas of a typical localizer.

The lower antenna is at exactly half the height of the upper antenna, and as may be seen from the patterns of Fig. 62, its vertical pattern will have a maximum value at the glide-slope position. This antenna, however, radiates carrier as well as the side bands. The 90-cps side-band frequency is in phase with the similar side band in the larger-angle lobe of the upper antenna and is in phase opposition with the similar side band in the smaller lobe of the upper antenna. This condition is reversed for the 150-cps side band. Accordingly, above the glide slope, a preponderance of 90-cps signal exists, while below the glide slope the preponderance is in favor of 150 cps. At the glide-slope angle, there is no contribution from the upper antenna, and since the 90- and 150-cps side bands of the lower

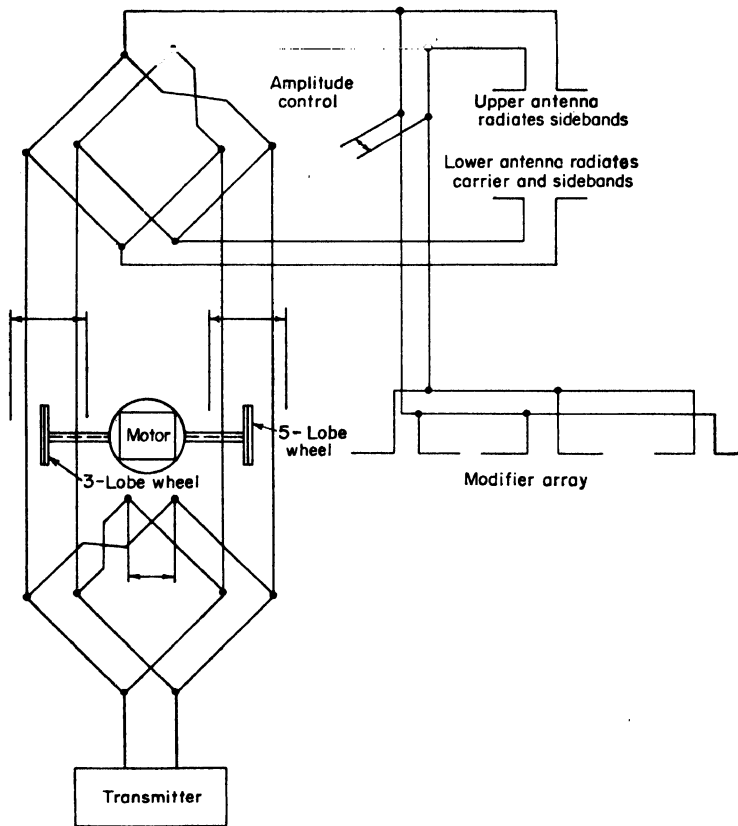



FIG. 61. Null-type glide-slope system.

antenna are equal, the horizontal needle of the cross-pointer instrument will indicate center. If 90 or 150 cps predominates a fly-down or fly-up indication will result, respectively.

The position of the slope in this arrangement is determined only by the position of the null that is generated by the upper antenna. Since the upper radiator is at a height of 33 ft for a 2-deg glide angle, snowfall as high as 2 ft would produce a change in path angle of only 0.13 deg. If the snowfall were as much as 4 ft, the glide slope of the equisignal system would be completely obliterated. However, such a snowfall would cause a change of less than 0.4 deg in the null-type system. It is also evident from the field patterns that the rate of change of indication for about 1.5 deg above and below the course is essentially uniform and leads to a linear deflection of the instrument.

The upper and lower antennas are similar in construction and consist of single dipoles located about $\lambda/4$ in front of reflecting screens.

As discussed for the equisignal-type glide slope, displacing the equipment to the side

of the runway produces a hyperbolic shape at the lower end of the slope. In the equisignal system, a straight-line slope was obtained by altering the relative intensity of the patterns in the horizontal plane. This method is not applicable to the null-type system since the slope position is determined only by the upper antenna. The third or modifier antenna is used for this purpose. The modifier antenna transmits (in the vicinity of the final let-down area or touch-down point) some additional side-band power in such a manner that the existing 90-cps side bands from the other antennas are augmented while the 150-cps signal is decreased. By properly shaping the horizontal pattern of the modifier antenna and controlling its level with respect to the antennas of the main array, the desired straight-line path is accomplished.

83. The Glide-slope Apparatus. Reception of the null-type signal is accomplished with the same glide-slope receiver that is used with the equisignal glide slope, and the ground equipment is so adjusted that it is unnecessary to change any receiver constants in order to fly either system interchangeably.

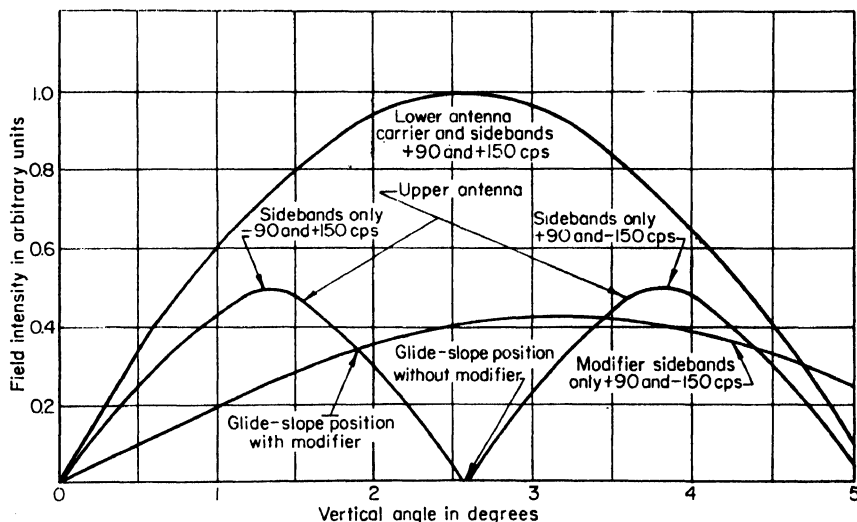



Fig. 62. Vertical pattern for the three antennas of the null-type glide-slope system.

As in the localizer equipment, it must be borne in mind that the modulation is accomplished through detuning; therefore the maximum power output of the transmitter must be equal to at least the peak power during the modulation cycle. Accordingly, good practice requires that the transmitter power capability be from seven to ten times the nominal carrier output required from the modulator terminals. It has been the practice, in the design of early transmitters, to provide for balanced output systems. Such transmitters employ crystal oscillators operating at about 6 Mc, followed by a series of conventional frequency multipliers to obtain the output frequency of 335 Mc. The final stage consisted of a grounded-grid amplifier and powers as much as 75 watts were obtained.

Modern transmitters employ a coaxial output and are modulated by a bridge modulator also employing coaxial transmission lines. The latter transmitters employ only 5 tubes and 5 adjustments as compared with 8 tubes and 9 adjustments for similar earlier transmitters. The power output of the modern transmitter is in excess of 100 watts.

Many air-borne glide-slope installations presently use a simple horizontal dipole mounted on a vertical standard at a height of about 12 in. above the aircraft fuselage. In one design, the horizontal dipole is mounted on the same vertical standard that supports the localizer antenna referred to as a *ram's horn*. More recently, however, the tendency has been to utilize flush-mounted antennas. These must be mounted in the forward part of the aircraft and have a curved configuration to conform to the shape of the nose of the aircraft. For this reason, although a large number of flush-mounted antennas are on the market, externally mounted antennas may be more popular. One of these antennas has a U shape

that projects horizontally from the nose of the aircraft. The plane of the U is horizontal and its maximum dimension is 5 in. Both ends of the U are terminated in insulating material near the fuselage.

During the Second World War, a glide-slope receiver in common use was the R89, later designated as the AN/ARN-5. These receivers provided reception on three channels in the band from 332.6 to 335.0 Mc. Modern glide-slope equipments provide 20 channels in the band from 329.3 to 335 Mc. A typical receiver weighs approximately 15 lb and utilizes 12 tubes.

One of the operational problems encountered with flying glide slopes is the convergence of the patterns at short distances to the equipment. This convergence causes the path to become extremely sharp and produces undue deflections of the cross-pointer indicator for slight deviations of the aircraft from the slope. The process of eliminating this effect has been termed *course softening*. In early systems, this was accomplished by an additional transmitter operating at the same nominal frequency as the main glide-path transmitter and radiating an unmodulated carrier from a directive array on the same mast. This array was directed toward the runway so that the received magnitude increased as the aircraft approached the touch-down point. The unmodulated carrier desensitized the receiver and thereby reduced the path sharpness as required. In modern systems, however, two general methods are incorporated in the receiver. One method is the application of the gain-control voltage to the audio stages. With such circuitry the relation of output to input may be given any desired shape depending upon the amount of control voltage that is applied to the section of the receiver that follows the detector. Another general method is to utilize an amplified AGC. Some modern receivers use both these methods to obtain the desired degree of course softening.

84. ILS Monitors. ICAO adopted the practice that suitable equipment should be located in the radiation field of the localizer and glide-slope equipments to provide signals for the operation of automatic monitoring equipment. The monitors are required to transmit a warning to a control point and to remove modulation or signal if any of the following conditions arise:

For the Localizer:

1. Shift of the on-course line by more than ⅓ deg from the center line of the runway
2. Reduction of power output to less than 50 per cent of the normal
3. Change of course width of more than 20 per cent

For the Glide Slope:

1. Shift of the path angle by more than 10 per cent of its chosen value
2. Reduction of power output to less than 50 per cent of normal
3. Increase in slope width of more than 10 per cent over the standard value

The above specifications are normally considered to be minimum and monitoring is generally accomplished to somewhat greater accuracies than those specified above.

85. Approach Control Using Heading and Rate. The localizer and glide-slope deflections of an ILS indicate lines of position rather than linear error from the course. At great distances from touch-down, a deflection on the localizer needle requires the pilot to change his heading. Because of the angular nature of the course, however, it will be some time before he crosses the desired course position. On the other hand, when close to the touch-down point, even a small change in heading may cause the aircraft to overshoot. The optimum approach is one which smoothly and asymptotically approaches the course line at any point during the approach. As a minimum, such an approach requires the use of a directional gyro, the bank indicator, and a knowledge of the rate of closure with the course. Circuits have been developed utilizing these factors which solve the problem of an asymptotic approach. One such circuit, developed by the Sperry Gyroscope Company, has been termed the *zero reader*.[85] The information is presented to the pilot on his localizer needle, and in this instance, a centered

needle indicates that the aircraft heading is momentarily correct to make the asymptotic approach rather than the fact that the pilot is centered on the course.

Equipment of this type is employed in automatic pilots for making approaches. In such equipment, the output of the heading-rate-localizer circuit is used to actuate servomechanisms which in turn actuate the aircraft controls. In fact, equipment is commercially available whereby both the localizer and glide-slope signals may be used for complete automatic approaches. This automatic technique[86] has been carried to a point where successful touch-downs have been made experimentally, but it is not yet considered adequate for routine use.

86. Radar Low-approach Systems. A radar low-approach system is intended to supply the same type of information as a fixed-beam low-approach system and the same general definition applies. In April, 1941, personnel assembled by the National Defense Research Council at the Radiation Laboratories associated with the Massachusetts Institute of Technology began to consider means for applying microwaves to the problem of low approaches. Two main goals were to be achieved by this equipment; namely, (1) to circumvent the sensitivity to siting that characterized beam systems developed up to that time, and (2) to make possible the all-weather landing of minimum equipped aircraft (for example, carrying communications only) or aircraft in which the navigational equipment had suffered damage. Flight tests, conducted with an antiaircraft gun laying radar, met with some success and pointed to the modifications required to optimize the equipment for low-visibility approaches. The development that followed led to an equipment called ground-controlled approach, or GCA. The equipment consisted of two radars, one for surveillance and guidance of the aircraft to an optimum approach position, and the other of a more accurate nature for the approach itself. Approximately 200 of these equipments were produced in the United States during the war by three manufacturers and were employed with considerable success by the Armed Services.

At a meeting of a Special Radio Technical Division of the PICAO in the fall of 1947,[87] the GCA equipment was recognized by the statement "that since GCA can perform certain very useful special functions, this system can be employed at those airdromes where local conditions warrant its use as an aid supplementary to the standard system."

The above statement reflected a growing realization by civil air operators of the usefulness of a radar low-approach system as an adjunct to the beam-type system to increase the over-all safety in low-approach operations. This opinion was seconded in the following statement[88] of the RTCA, Special Committee 31, which met between June 12, 1947, and Feb. 17, 1948: "It was agreed that ILS and Precision Beam Radar, when used by aircraft equipped to do so, are complementary and when used together furnish the most dependable landing information." The report included a recommendation for a program to install precision-beam radar equipment at suitable airports where operational requirements justified this action. This decision led to a similar recommendation[89] by the ICAO in its meeting in 1949.

Many airports throughout the world are presently equipped with GCA equipment and, in most instances, the equipment is utilized in support of the beam-type ILS as a very effective monitor and stand-by equipment. There are many instances, however, where the GCA alone has provided successful all-weather approach information.

87. Principles of Radar Low-approach System. GCA consists of two radars termed the *surveillance radar element* (SRE) and the *precision-approach radar* (PAR). The SRE or search system is used for locating all aircraft in the vicinity and directing them to the correct position for landing. The general principles of operation are similar to those described for the surveillance radar in Sec. 75. The SRE component of GCA utilizes a frequency of 3,000 Mc, and the wartime units utilized an r-f peak power of 80 kw. The p.r.f. is 2,000 pulses per sec with a pulse duration of 0.5 μsec. The antenna utilizes a (cosecant)^2 pattern in the vertical plane so as to eliminate a large proportion of ground clutter. The rotation or scan rate is 30 rpm. The PPI display is normally superimposed on a map of the surrounding area so that the operator is able

to tell the exact location of any particular aircraft and, accordingly, can instruct the pilot to avoid obstacles in his path.

Strictly speaking, only the PAR component is used for the low approach. As explained for beam systems, the approach path is defined by the intersection of two planes, which in ILS constitute the localizer and glide slope. A similarity exists in GCA in that two highly directive beams are scanned in the vertical plane and horizontal plane, respectively, and the positional information derived from these defines the course to be followed.

The azimuth antenna (scanning horizontally) utilizes a pattern 0.6 deg wide in azimuth and 1.5 deg wide in elevation. The elevation antenna (scanning vertically) utilizes a pattern 0.4 deg wide in elevation and 3 deg wide in azimuth. The elevation antenna scans 7 deg in the vertical plane while the azimuth antenna scans 20 deg in the horizontal plane. Because of the high directivity required, the antenna apertures must be many wavelengths; hence microwave frequencies (10,000 Mc) are used to keep the antenna structures relatively small. One of the commercial antennas for scanning in the horizontal plane has a length of approximately 8.5 ft and employs 128 dipoles. A typical array for scanning in the vertical plane is approximately 14 ft in length and contains 168 dipoles. Both the azimuth and elevation antennas scan electrically, two to four times per second, by means of a system which mechanically varies the width of the wave guide feeding the transmitting dipole stacks. The mechanism has been termed the *squeezable wave-guide antenna* and was developed by Dr. L. W. Alvarez. Briefly, this mechanism involves the following principle. The dipoles of the array are fed by probes inserted in a common wave guide which runs the length of the array. The individual amplitude is controlled by adjusting the length of the probe, and the initial phase conditions are established by the spacing of the dipoles. By changing the cutoff wavelength of the wave guide, it is possible to change the phase velocity and therefore the relative phase of the excitation delivered to the individual elements of the array. This is accomplished by means of a motor-driven scissors-type mechanism which varies the cross section of the wave guide in the required cyclical manner.

A typical PAR transmitter utilizes a p.r.f. of 2,000 pulses per sec and a pulse duration of 0.5 μsec with a carrier frequency of 10,000 Mc. The energy is applied alternately to the azimuth and elevation antennas by a motor-driven r-f switch. The scanning mechanism is geared to the same shaft and is thus synchronized with the operation of the switch. It further supplies a commutator signal so that the output of the receiver appears alternately on the azimuth and elevation indicators. After careful alignment of the PAR with the runway, the exact position of an aircraft on approach is determined by the beam positions at the instant the echo is received. Since the energy is pulsed, conventional timing techniques are utilized to derive the distance of the aircraft.

88. Instrumentation and Accuracy of the Radar Low-approach System. The effectiveness of the radar approach system is closely associated with the manner in which the data are displayed. During the war, the SRE of the GCA was displayed on a PPI employing range scales of 7.5, 15, and 30 miles. The PAR was displayed on two cathode-ray displays for the elevation and azimuth components, respectively. Each precision indicator utilized two cathode-ray tubes, one covering a 10-mile range (which is the limit of the precision system) and the other a 2-mile range. Typical displays of this type are shown in Fig. 63. To utilize this display, the operators manipulated controls that positioned illuminated lines located on transparent plastic scales so that they were over the spots representing the target. Attached to the scales were cam mechanisms that rotated voltage dividers. The cams were cut with curves that corresponded to the localizer and glide-slope positions that an aircraft should follow in making an optimum approach. With an aircraft making such an approach, the cam positioned the arm on the voltage divider so that the output voltage was zero. A deviation above or below, or to the right or left of the optimum path, produced a corresponding positive or negative output voltage. This voltage was connected to a cross-pointer instrument similar to that used for the ILS beam approach. This instrument was viewed by a third operator who used its indication for calling instructions to the aircraft via the communication equipment.

In civil operations, three operators was considered extravagant, and this requirement led to the development of the presently used display known as the *az-el scope,* on

which there is displayed azimuth, distance, and elevation information. With this equipment, a single operator can call instructions directly from the az-el scope. The display is shown in Fig. 64. Since the azimuth and elevation antennas are alternately excited, the upper, or elevation, display and the lower, or azimuth, display are produced at different times in synchronism with the antenna switching. Both displays are of the expanded type and have their origins near one edge of the cathode-ray screen. For the elevation display, approximately 6 deg of tube surface is equal to 1 deg of elevation, while 3 deg of tube surface is equal to 1 deg in azimuth.

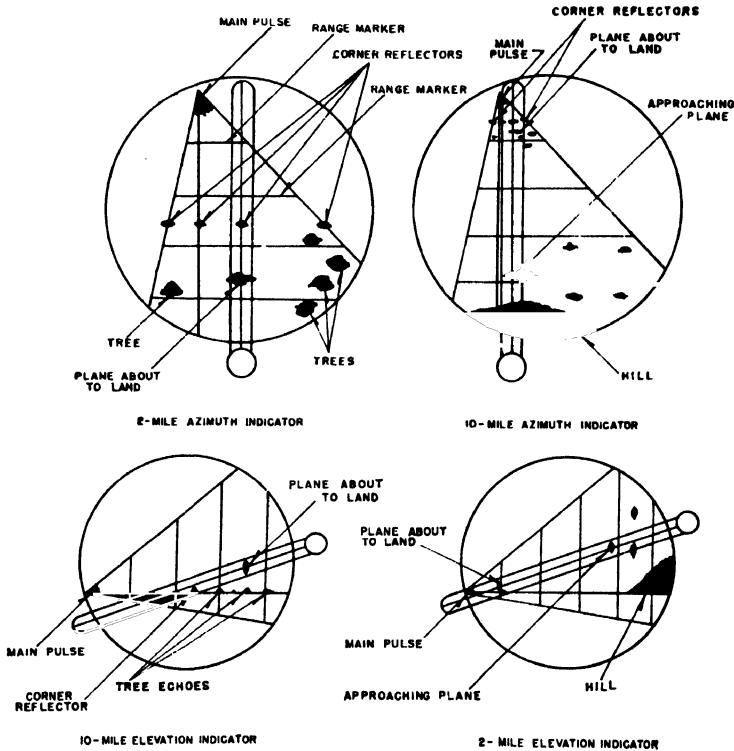



FIG. 63. Information displayed to operators of precision indicators of GCA "talk-down" radar landing system.

Through electronic means, range marks are reproduced on both displays. The total range of the display is 10 miles, but the calibration is nonlinear so that about one-half the display is equal to 3 miles while the remaining half is equal to 7 miles. The range marks incorporate gaps which serve to indicate the direction in which the scanning antennas are pointing. In modern equipment, the operator may adjust the direction of the antenna so that the sweep takes place about this average position. The indication of direction on the scope therefore reminds the operator of the setting of the antennas.

The optimum descent and azimuthal courses are also reproduced on the face of the tube through electronic means. These courses are straight lines in space, but because of the expansion of scales, they appear curved on the face of the display. The circuitry required to produce the az-el display is relatively extensive and includes as its main components antenna-position voltage sources, target-position indicating circuits, chart-generating circuits, cursor-generating circuits, and range-mark generators.

The power output of the transmitter used in the PAR element need not be great because of the high order of gain obtained with the antenna systems. Early designs employ peak powers of only 15 kw, while newer designs have employed powers of 45 kw. Pulses are short, having a duration of approximately 0.2 μsec and rise times of 0.1 μsec.

MTI circuits are employed in some modern equipment, thereby necessitating high pulse-repetition rates of the order of 5,500 per sec. When MTI is not employed, the pulse-repetition rate is normally reduced to about 2,000 per sec.

Modern equipment is capable of resolving targets that are separated by 200 ft in distance, 1.1 deg in azimuth, and 0.6 deg in elevation. At a distance of 1 mile, it is possible to detect deviations from the glide slope of as little as 25 ft. Common twin-engined aircraft can be

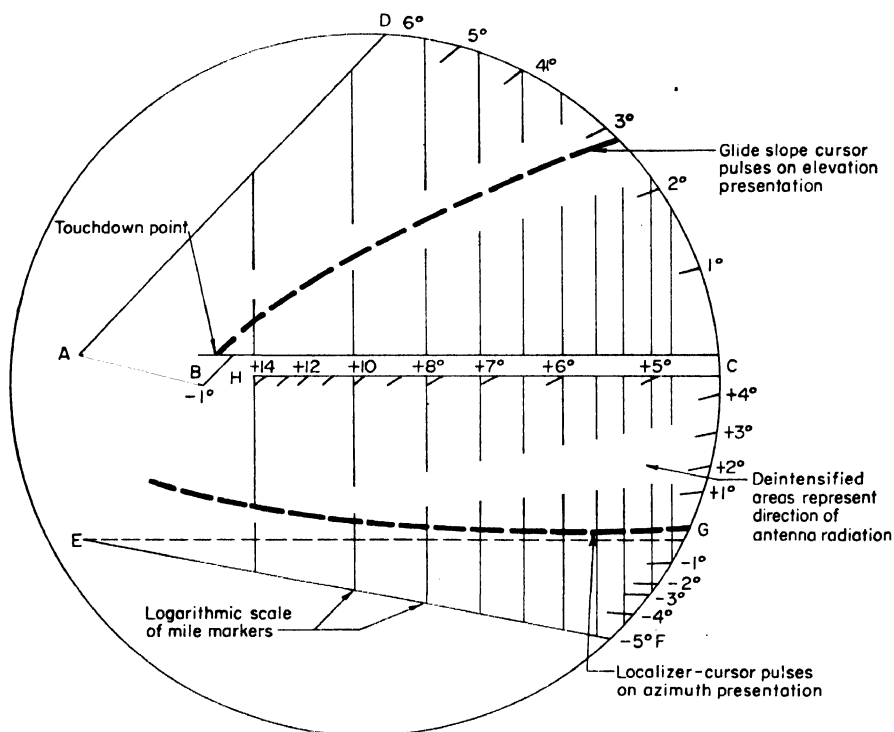



FIG. 64. Azimuth-elevation display of postwar low-approach radar.

detected at a distance of about 8 nautical miles. The bearing of an aircraft from the radar can be determined with an accuracy that is equal to 0.6 per cent of the range plus 10 per cent of the deviation from the optimum approach path. In elevation, an aircraft may be detected with an accuracy equal to 0.3 per cent of the range plus 10 per cent of the deviation from the glide slope. Distance to the aircraft may be determined within an accuracy equal to 2 per cent of the distance, provided the aircraft is at a distance of more than 5 per cent of the sweep range in use. The above accuracy figures are based on the assumption that the equipment is operating under conditions of accurate calibration and do not include possible errors due to drift in calibration with time.

89. Landing Altimetry. During the operational use of ILS and GCA equipment during the Second World War, it became evident that no line-of-position guidance system, regardless of its accuracy, was adequate to permit consistently safe landings to the touch-down point in zero visibility. It is not sufficient that the aircraft be

guided to a certain point on the runway, but a landing system must in addition provide a safe rate of descent at the touch-down point and ensure proper orientation of the aircraft. Contrary to electronic systems, during a visual landing the pilot does not attempt to fly a specific path during the last few hundred feet of the landing configuration. If the aircraft is blown to one side or the other (or even vertically) the pilot normally assumes a new path and flare-out condition. The primary concern is the distance above ground, and the rate at which the aircraft is approaching ground. The problem, therefore, subdivides itself into an approach phase and a landing phase. After an approach system similar to those described in Part 3 has brought the aircraft to an accurate position at a height from 50 to 200 ft, a visual or electronic means must be provided whereby the height and rate of descent become accurately known. The electronic means is known as an *absolute altimeter* or *terrain-clearance indicator.*

While the need for supplementing a good approach system with absolute altimetry became evident during the Second World War, the importance of altimetry to air operation was realized as early as 1924 when Alexander Behm[90] invented a sonic altimeter. Other altimeters[91] employing sonic, capacitive, and radio principles were developed in the years from 1924 to 1938. The first quasi-commercial device, however, was produced by the Western Electric Company in 1938 at the request of United Airlines. During the war years, the Radio Corporation of America produced a large number of radio altimeters similar in principle to the Western Electric device and identified as AYB-1 and RC-24. Following the war, many manufacturers entered the field of radio altimetry, and one of the outstanding units, known as the STR-30A, was developed by Standard Telephone and Cables, Limited, London, England.

Operating altimeters have been constructed using the above-mentioned sonic, capacitive, induction, and radio techniques. Although limited success was attained, the sonic, capacitive, and induction techniques are not popular in modern applications and exhibit many practical difficulties. By far the most widely used and commercially successful models are those employing f-m techniques, and accordingly the description that follows is limited to this type.

90. Principles of Frequency-modulated Radio Altimeter. In the pulse altimeter described in Part 1, the minimum distance that can be determined accurately is a function of the duration of the pulse. For example, with a pulse duration of only ¼ μsec the minimum distance would be approximately 250 ft. For landing applications, the accuracy must be considerably better; hence, the pulse-type altimeter is clearly unsuitable for this purpose.

In the f-m altimeter, the frequency of the transmitted wave is varied at some cyclic rate. In the original Western Electric and RCA altimeters, the frequency was varied from its lowest to its highest value in 1/240 sec, but postwar altimeters utilized sweep frequencies of four times this rate.

In operation, a transmitter output is connected to a suitable antenna which directs its energy toward the ground. The energy reflected by the ground is received on a second antenna connected to a radio receiver. The block diagram of a basic system is shown in Fig. 65. Some of the energy from the transmitter is also received by the receiving antenna; hence a beat between the direct and reflected wave appears at the receiver input. The total frequency change between the received and transmitted wave is obviously the rate of change of frequency multiplied by the time required for the transmitted wave to return. Since, as shown in Fig. 66, the rate of change of frequency is maintained constant, the beat frequency will be proportional to the reflection round-trip time. Hence, the beat-note frequency becomes a direct indication of altitude.

The minimum distance at which frequency may be read is a function of the rate of change of frequency and the frequency deviation. The RCA altimeter, operating at 440 Mc, incorporates a deviation of 40 Mc occurring 120 times per sec. With these parameters, the minimum usable distance becomes 19.5 ft. By utilizing a frequency-measuring system that will read to an accuracy of 10 cycles, distance can be read to approximately ±½ ft.

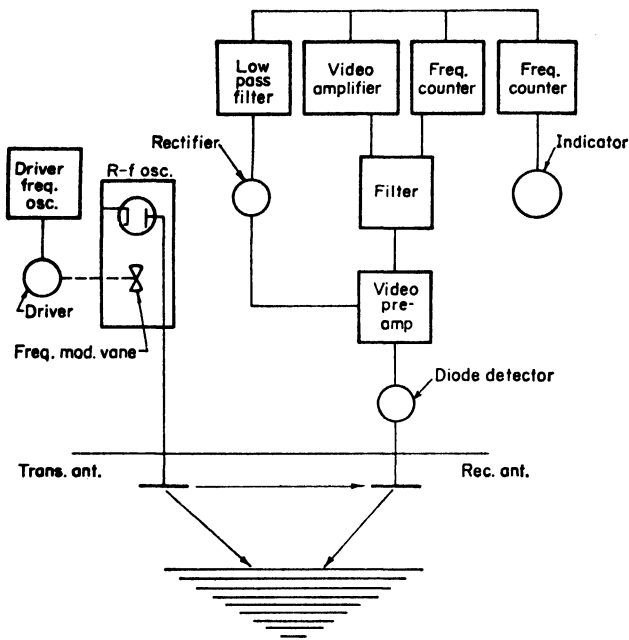



FIG. 65. Frequency-modulated radio altimeter system.

The Western Electric altimeter operated in the range from 420 to 445 Mc, and the RCA altimeter operated in the range from 420 to 460 Mc. Therefore, the transmitters employed in these equipments were relatively similar. The power output of the Western Electric altimeter was 10 watts, but by the time the RCA altimeter was designed there had been a marked improvement in receiver techniques; therefore, its transmitter was required to produce a power output of only 0.1 watt.

The postwar altimeters have been allocated the band from 4,200 to 4,400 Mc. One of these altimeter transmitters employs a small magnetron while the STC transmitter employs a form of klystron. In the Western Electric altimeter, frequency was varied by a motor-driven capacitor. The STC device followed the same practice. In one American postwar altimeter, as well as in the older RCA device, modulation of the transmitter frequency was accomplished through the use of vibrating reeds located so as to form one side of a capacitor in the oscillator circuit.

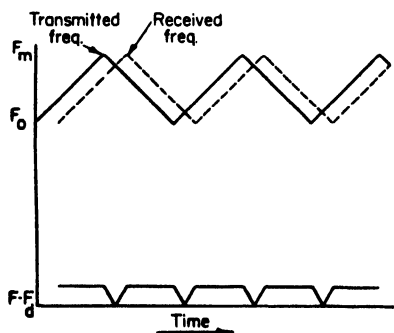



FIG. 66. Transmitted and received frequency of radio altimeter together with generated beat note. The beat note disappears at various times during the frequency sweep cycle.

Both the Western Electric and RCA altimeters made use of a form of folded dipole for antennas. The dipoles were about $\lambda/2$ in length, which corresponds to approximately 12 in. The dipoles were horizontal and supported by vertical structures at a distance of about $\lambda/4$ beneath the fuselage of the aircraft. The postwar altimeters

(operating in the 4,000-Mc band) have employed electromagnetic horn antennas. In one American design, the antennas form part of the receiver-transmitter assembly and therefore do not require additional cabling. The horn openings are normally covered with a Fiberglas plate and mount flush with the skin of the aircraft so as to avoid aerodynamic drag.

The indicator circuits of the different altimeters vary in detail. Essentially, however, all consist of a high-gain audio amplifier, an automatic frequency-response control circuit, and a counter circuit. The counter circuit measures cycles per second and indicates the altitude on a meter calibrated in feet. The altimeters from the three major manufacturers are designed to give height to 5,000 ft and include low-altitude scales. As an example, the STC design incorporates a linear scale reading from 0 to 5,000 ft. A second instrument is used for lower altitudes which has an exponential scale so that 0 to 100 ft is indicated in about 135 deg while the remaining 135 deg reads from 100 to 500 ft.

It has been indicated previously that rate of descent is a vital factor in the safe landing of aircraft. In addition, for automatic operation, the second derivative of the altitude (or vertical acceleration) must also be known for proper stabilization of the automatic pilot equipment. For velocity determination, a circuit is used in which a reference pulse samples the received signals, the phase change being converted to a low frequency which is proportional to velocity. This frequency is connected to a counter similar to that used in altitude determination, and furnishes a d-c output proportional to rate of descent or climb. Differentiation of this quantity provides vertical acceleration.

For low-visibility automatic-landing service, the output of the altimeter is combined with the glide-slope signal. Because the rate information increases in importance at the lower altitudes, a circuit is normally employed which deemphasizes the effect of the glide-slope information as the ground is approached. The glide-slope information, on the other hand, swamps out the rate and height information as height increases. The combined information is applied to the horizontal needle of the cross-pointer instrument, which now dictates the optimum landing path rather than merely the location of the glide slope. Systems of this type have been successfully flown experimentally but are not yet considered adequate for routine commercial operation. One of the major problems is the convolutions of the ground in the vicinity of the approach runway. Unless the ground is relatively flat for several hundred feet before the paved runway begins, altimeter aberrations will result during the critical landing phase. This problem is receiving the serious consideration of designers of instrument-landing equipment.

PART 4

THE EQUIPMENT OF THE AIRPORT ZONE (SURFACE DETECTION)

The airport zone has been defined[92] as "that area within the boundary of an airport primarily used for the movement of aircraft and other airport vehicles between various points on the airport excepting (a) the runway or runways currently being used for takeoff and landing of aircraft and (b) localized areas primarily used for the storage and maintenance of aircraft when movement within such areas does not constitute a collision hazard with airport traffic." The importance of the airport zone and its problem was recognized by the Special Radio Technical Division of the PICAO, which, in its meeting of the fall of 1946, expressed itself as follows:[93] "The Division notes the importance of the movement area (airport zone) as being a limiting factor in high density traffic areas. Considerable work has been completed to increase the rate of landing by various means, but the taxiing and dispersal problem has, in comparison, been neglected. The capacity of an airdrome must depend ultimately upon the rate of clearing the runway in use."

91. Requirements. The complexity of taxiways at large modern airports makes it mandatory that control be vested in a central controller who has an over-all view of all vehicles on the airport. When instrument weather prevails, it may not be possible for the controller to observe more than a fraction of the total surface.[94] Accordingly, the following requirements for the airport surface-control system have been formulated:

1. The system must provide a positive indication to the controller of the location of all vehicles on the surface of the airport.
2. The system must be capable of providing the pilot of an aircraft with accurate indication of the route to be followed between the starting point and ultimate destination.
3. The system must permit the selection of any of a series of routes from one point on the airport to another.
4. The system must provide automatic control of traffic and taxiway intersections in such a manner that access to these intersections will be permitted or denied as dictated by collision hazards.
5. The system shall provide safety checks to prevent aircraft collision.
6. The system must report back to the controller the progress which a vehicle has made along the selected path.
7. The system must automatically signal to the controller any dangerous situation arising from the interpretation of control information.
8. The system must provide a means for passing of control from one form to another, that is, from landing control to ground-movement control.

The above requirements indicate that three general types of equipment are needed, as follows:

1. Equipment for determining the position of all vehicles
2. Equipment for providing information as to the identity of all vehicles
3. Equipment for providing guidance instructions

At present, the equipment for item 3 above consists of normal communications gear. However, the advantages of a visual presentation are evident, and serious consideration is being given to this problem. In lieu of item 2, aircraft may be identified through the knowledge of the order in which they are fed into the traffic pattern or from knowledge of their order of departure. Control panels linking traffic to airport control have been devised for this purpose. Item 1 covers a series of devices known as "airport-surface-detection equipment," and these are described in the sections to follow.

As of early 1957, ICAO had not adopted a standard airport-surface-detection (ASD) equipment. In its meeting[95] of 1946, various types were considered, including the pad system with pressure-actuated switches and with light-beam switches, the magnetic-loop system, and the radar system. None of these systems was considered to meet all the desired functional requirements. In recent years, considerable additional work has been done but the actual installation and use of ASD equipment for routine airport surface movement has lagged far behind that for the en-route long- and short-distance zone and the approach and landing zone.

92. Traffic-pad Equipment. The ASD equipment of the traffic-pad type is intended to supply to a controller a miniature display of the entire airport on which there is indicated the position of all vehicles. Traffic pads of an early type were tried at the La Guardia Airport in New York in the years between 1948 and 1950. These were of a simple type incorporating a rubber hose which was placed across the taxiway and which actuated a pressure switch when compressed by a passing vehicle. The switches flashed signals in the tower which were momentary in nature and therefore did not indicate a vehicle stopped between two such paths. Another disadvantage of the system was hindrance to the operation of snow-removal equipment. Proposed systems of a preferred type utilize metal pads level with the surface of the taxiways or detecting devices utilizing the infrared spectrum. Whatever the system for the detection of a passing aircraft, the resulting momentary signals must be associated with other circuits in order to present a continuous display of position.

An example of a proposed block system is one in which the detection equipment is installed in pairs. The distance between elements of the pairs is approximately 40 in.,

with the next pair of elements occurring approximately 150 ft for the length of taxiway to be controlled. Each 150-ft section between pairs then constitutes a movement block. The dimensions given are typical, but any dimensions can be used which experience shows to be desirable. The pairs are connected to relay and indicating circuits which incorporate memory devices such as holding relays or rotary switches. The direction of the vehicle is determined by the sequence of the triggering of the pairs, and the block is released only when the succeeding pairs have been triggered. Circuits are also incorporated for indicating the total number of vehicles that have entered any given block. This provision permits a repair vehicle to enter an occupied block to service an incapacitated aircraft. The information is displayed on a chart showing the occupancy of traffic-movement blocks and the number of vehicles within any block. If it is desired to indicate to the aircraft that the block ahead of it is occupied, a system of traffic lights may be used along the edge of the taxiways.

93. Magnetic-loop System. A magnetic-loop system has been installed and evaluated at the London Airport.[96] This system consists of cables installed in the bitumen joints that provide for expansion between concrete sections. The installation is shown diagrammatically in Fig. 67. In this system, loops of wire are wound around two taxiway sections so that two coplanar coils have a 50 per cent overlap. When excited by an oscillator this arrangement provides essentially zero coupling between adjacent loops.

It should be noted that the pair of coils forms two arms of a bridge circuit, the bridge being completed by additional adjustable elements. The oscillator operates at an audio frequency. Without an aircraft present, the bridge is adjusted for balance and no output results. When an aircraft enters the loop area, coupling between the loops assumes some finite value. In addition, the inductance of the individual loops decreases and the resistance increases. The net effect is an appreciable increase in the potential drop in one arm of the bridge and a decrease in the opposite arm. The result is an output from the bridge (and associated amplifier) which actuates lamps on a chart of the airport. Circuits are also employed to operate lights which are mounted flush with the surface of the taxiway and serve to warn a pilot that the block ahead is occupied.

FIG. 67. Schematic of magnetic-loop system.

94. Airport-surface-detection Radar. An attractive feature of airport-surface-detection radar is its ease of installation. It is recognized that radar serving such a purpose must be capable of very high target resolution; accordingly, development was conducted at wavelengths of 1 cm in England during the late war years and similar work was undertaken in the United States in about 1947 and resulted in a satisfactory test model.[97]

The ASD radar operates at a wavelength of 1.25 cm and has a peak power of 20 kw. Definition is obtained by a relatively large antenna consisting of a horn-fed truncated parabola having a width of 7.5 ft and a height of 3 ft. The antenna produces a field pattern which has a width in the horizontal plane of 0.4 deg measured between half-power points. The pattern in the vertical plane approximates a (cosecant)^2 function and is approximately 10 deg high. The antenna rotates at a speed of between 30 and

60 rpm and is mounted on a tower approximately 30 ft high. Pulse lengths of between 0.015 and 0.08 μsec are available, the most commonly used being 0.02 μsec. The indication is presented on a 12-in. PPI display.

The radar is capable of detecting utility vehicles such as a jeep or fire truck. A typical display is shown in Fig. 68. A twin-engine Douglas DC-3 aircraft is plainly visible on the taxiway. The range accuracy was determined by test to be well under 5 ft. With targets located at a distance of 1,000 ft, it was found possible to resolve two targets when they were separated by 35 ft. At distances of 5,000 ft, it was necessary that the targets be separated by 65 or 70 ft.

A major disadvantage of the ASD radar is that the extremely high frequencies (1 cm) suffer bad propagation attenuations in certain types of precipitation, thereby making the device unreliable even for the short ranges required in this application.

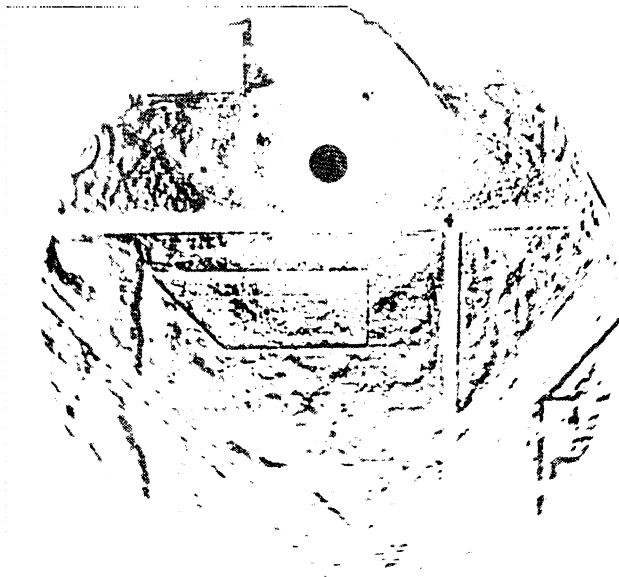



FIG. 68. Display on United States airport-surface-detection radar. Note aircraft. (*Courtesy of CAA.*)

It was found that in light rainfall (0.08 in. per hr) in which the visibility had dropped to less than 1 mile, little effect could be noticed in the operation of the radar. With moderate rainfall (0.25 in. per hr) which produced visibilities of about ½ mile, a Douglas DC-3 could still plainly be discerned on the radar. With rainfall of the order of 0.4 in. per hr, classified as heavy rainfall, the Douglas DC-3 was bearly discernible on the radar.

It was concluded that the ASD radar was satisfactory to provide sufficient information to pilots on runways or wide taxi strips and to provide safe separation between converging aircraft operating on the airport. However, the resolution was not considered adequate for control of aircraft on the terminal ramps or in congested parking areas.

References

1. Radio Technical Commission for Aeronautics: "Recommended United States Policy, Air Navigation-Communication-Traffic Control," Aug. 28, 1946 (corrected Oct. 30, 1946), p. 29.
2. Gringorten, I. I.: Minimal Flight Paths, *J. Inst. Nav.*, 1 (9), 194, 1948.
3. Special Radio Technical Division, Provisional International Civil Aviation Organization: "Final Report—First Session," January, 1947, pp. 25-43, Montreal, Canada.

4. Communications Division, International Civil Aviation Organization: "Report of the Fifth Session," June 4, 1954, Com V-WP/171, pp. 111-113, Montreal, Canada.
5. Kruesi, G. G.: "Radio Direction Finder," U.S. Patent 1868945, July 26, 1932.
6. Moseley, F. L.: "Radio Compass Navigation Apparatus," U.S. Patent 2257757, Oct. 1, 1941.
7. Harrison, A. E.: Geographical Separation of Radio Range Stations Operating on the Same or Adjacent Frequencies in the 200-400 Kc. Band, *CAA Tech. Development Rept.* 4, January, 1938.
8. Shangraw, C. C.: Radio Beacons for Transpacific Flights, *Proc. IRE*, September, 1928, p. 1203.
9. Special Radio Technical Division, Provisional International Aviation Organization: "Final Report —First Session," January, 1947, pp. 25, 43, Montreal, Canada.
10. Adams, P. R., and R. I. Colin: Frequency, Power, and Modulation for a Long-range Radio Navigation System, *Elec. Commun.*, **23**, 145, June, 1946.
11. Sreenivasen, K.: On the Relation between Long-wave Reception and Certain Terrestrial and Solar Phenomena, *Proc. IRE*, **17**, 1793-1813, October, 1929.
12. Schelleng, J. C.: Some Problems in Short-wave Telephone Transmission, *Proc. IRE*, **18**, 913-938, June, 1930.
13. Anderson, C. N.: North Atlantic Ship-shore Radiotelephone Transmission during 1932-1933, *Proc. IRE*, **22**, 1215-1224, October, 1934.
14. Eckersley, T. L.: An Investigation of Short Waves, *J. IEE*, **67**, 992-1029, August, 1929.
15. Beverage, H. H., and H. O. Peterson: Diversity Receiving System of R.C.A. Communications, Inc., for Radiotelegraphy, *Proc. IRE*, **19**, 531-561, April, 1931.
16. Magnetic Storm and Aurora of October 14-15 and Associated Solar Activity, *Nature*, **118**, 679-680, Nov. 6, 1926.
17. Wonnell, T. S., and G. E. Fenimore: C.A.A. Low Frequency Omni-radio Range, *CAA Tech. Development Rept.* 72, August, 1947.
18. Mitchell, H. T., and T. Kilvington: Time-multiplex Radio-frequency Phase-comparison Method for Navigational Systems, *J. IEE*, Part 3A, **94** (16), 977, 1947.
19. Clark, C. T., R. I. Colin, M. Dishal, I. Gordy, and M. Rogoff: Navaglobe-Navarho, Long-range Navigation System, *IRE Convention Record*, Part 1, 1954.
20. Busignies, H., P. R. Adams, and R. I. Colin: Aerial Navigation and Traffic Control with Navaglobe, Navar, Navaglide, and Navascreen, *Elec. Commun.*, **23**, 115, June, 1946.
21. Communication Division, International Civil Aviation Organization: "Report of the Fifth Session," Com V-WP/171, pp. 111-113, June, 1954, Montreal, Canada.
22. "The Decca Navigator—Air," Decca Navigator Company, Ltd., London, England, 1946.
23. "Delrac, A Long Range V.L.F. Area Coverage Navigational Aid," Memo Plans 154, Issue 4, Decca Navigator Company, Ltd., London, England, January, 1954.
24. "Dectra System for Long Range Air Navigation," Decca Navigator Company, Ltd., London, England, February, 1954.
25. Pierce, J. A.: Radux, *Cruft Laboratory Tech. Rept.* 17, July 11, 1947.
26. Adcock, F.: "Improvement in Means of Determining the Direction of a Distant Source of Electromagnetic Radiation," British Patent 130490, 1919.
27. Eckersley, T. L.: Scattering, Polarization Errors, and the Accuracy of Short Wave Direction Finding, *Marconi Rev.*, **53**, 1-8; **54**, 20-30, 1935.
28. Stansfield, R. G.: Statistical Theory of D. F. Fixing, *J. IEE*, Part 3A, **94**, 762, 1947.
29. Ross, W.: Estimation of the Probable Accuracy of High-frequency Radio Direction-finding Bearings, *J. IEE*, Part 3A, **94**, 722, 1947.
30. Barfield, R. H.: Statistical Plotting Methods for Radio Direction-finding, *J. IEE*, Part 3A, **94**, 673, 1947.
31. Pierce, J. A.: An Introduction to Loran, *Proc. IRE*, **34**, 216-234, May, 1946.
32. The Loran System, Part I, *Electronics*, November, 1945, pp. 94-99.
33. Communication Division, International Civil Aviation Organization: "Report of the Fifth Session," Com. V-WP/171, p. III-12, April, 1954, Montreal, Canada.
34. "Air Navigation," T.M. 1-205, p. 3, War Department, Washington, D.C.
35. Borden, R. C., and E. C. Williams: Radar Mapping of the Chicago-New York Airway, *CAA Tech. Development Rept.* 66, Technical Development and Evaluation Center, Indianapolis, Ind., April, 1950, p. 3.
36. Standards on Radio Aids to Navigation, Definition of Terms, 1949, *Proc. IRE*, **36**, 1365, December, 1949.
37. Hall, J. S.: "Radar Aids to Navigation," pp. 110, 135, McGraw-Hill, 1947.
38. Radio Technical Commission for Aeronautics: "Recommended United States Policy, Air Navigation-Communication-Traffic Control," Aug. 28, 1946 (corrected Oct. 30, 1946), p. 21.
39. Lyman, J., and G. B. Litchford: Effect of Position Determination on Air Traffic Control, *J. Inst. Nav.*, **2** (8), 1950.
40. Englund, C. R.: "Radio Direction Finder," U.S. Patent 1815246, July, 1931.
41. Evans, P. H., and J. W. Greig: "Radio Direction Finding," U.S. Patent 1933248, October, 1933.
42. Greig, J. W.: "Direction Finding System," U.S. Patent 1988006, January, 1935.
43. Luck, D. G. C.: Omnidirectional Radio-range System, *RCA Rev.*, **6** (1), January, 1941.
44. Luck, D. G. C.: Omnidirectional Radio-range System, *RCA Rev.*, **6** (3), January, 1942.
45. Stuart, D. M.: Omnidirectional Range, *Aero Digest*, June, 1945.
46. Alford, A., and A. G. Kandoian: Ultra-high-frequency Loop Antennae, *Elec. Commun.*, **18**, 255, April, 1940.
47. Lundburg, F. J., and F. X. Bucher: Cage-type Very-high-frequency Phase-comparison Omnidirectional Radio Range Antenna, *Elec. Commun.*, **29**, 108-116, June, 1952.

48. Montgomery, B. E.: Very-high Frequency Aircraft Antenna for the Reception of 109-megacycle Localizer Signals, *Proc. IRE*, **33**, 767, November, 1945.
49. Granger, J. V. N.: Designing Flush Antennas for High-speed Aircraft, *Electronics*, March, 1954, p. 140.
50. Air Technical Service Committee: "Report of Electronic Subdivision Advisory Group on Air Navigation," Wright Field, Dayton, Ohio, February, 1946, p. 32.
51. Special Radio Technical Division, Provisional International Civil Aviation Organization: "Final Report—First Session," January, 1947, p. 39.
52. Simson, L. H.: National Allocation Plan for Assigning Radio-range Frequencies in the Band 119–126 Megacycles, *CAA Tech. Development Rept.* 28, May, 1941.
53. Hirsch, C. J.: Pulse-multiplex System for Distance-measuring Equipment, *Proc. IRE*, **37**, 1236, November, 1949.
54. Dodington, S. H.: Crystal Control at 1000 Megacycles for Aerial Navigation, *Elec. Commun.*, **26**, 275, December, 1949.
55. Smith-Rose, R. L.: Speed of Radio Waves and Its Importance in Some Applications, *Proc. IRE*, **38**, 16–20, January, 1950.
56. Gross, F. J.: Course-line Computer for Radio Navigation of Aircraft, *Proc. IRE*, **37**, 832, July, 1949.
57. Gross, F. J., and Hugh A. Kay: Initial Flight Tests and Theory of an Experimental Parallel Course Computer, *CAA Tech. Development Rept.* **83**, September, 1948, p. 10.
58. Dippy, R. J.: Gee: A Radio Navigational Aid, *J. IEE*, Part 3A, **93** (1–4), 471, 1946.
59. "New Gee Airborne Equipment for Civil Aviation," Cossor Radar Ltd., Highbury, London (undated).
60. Development of the Ultra-high Frequency Radio Range, Part 3, *CAA Tech. Development Rept.* 49, June, 1945, p. 2.
61. "Statistical Handbook of Civil Aviation," p. 19, Government Printing Office, 1949.
62. "Standards and Recommended Practices," Aero. Telecomm. Annex 10, 2d ed., ICAO, Oct. 1, 1951, pp. 96–98.
63. RTCA Special Committee 31: "Air Traffic Control," Paper 27-48/DO-12, p. 23, May 12, 1948.
64. Busignies, H., and M. Dishal: Some Relations between Speed of Indications, Bandwidth, and Signal-to-random-noise Ratio in Radio Navigation and Direction Finding, *Proc. IRE*, **37**, 479, May, 1949. Also *Elec. Commun.*, **26**, 228–242, September, 1949.
65. Shannon, C. E.: Mathematical Theory of Communications, *Bell System Tech. J.*, **28**, 379–423, July, 1948.
66. Shannon, C. E.: Recent Developments in Communications Theory, *Electronics*, April, 1950, p. 82.
67. Busignies, H., P. R. Adams, and R. I. Colin: Aerial Navigation and Traffic Control with Navaglobe, Navar, Navaglide, and Navascreen, *Elec. Commun.*, **23**, 121–135, June, 1946.
68. Special Radio Technical Division, Provisional International Civil Aviation Organization: "Final Report, First Session," January, 1947, p. 15, Montreal, Canada.
69. Whiteley, E. A.: Spacing of Aircraft under High Density Approach Conditions, *J. Roy. Aeronaut. Soc.*, August, 1952, p. 581.
70. "Air Traffic Control and the National Security," Air Coordinating Committee, Air Traffic Control and Nav. Panel, Oper. Policy Group, December, 1950, p. 89.
71. Special Radio Technical Division, Provisional International Civil Aviation Organization: "Final Report, First Session," January, 1947, p. 22, Montreal, Canada.
72. Communications Division, Air Navigation Commission, International Civil Aviation Organization: "Report of Fourth Session," Apr. 24–June 2, 1951, p. 222, Montreal, Canada.
73. Radio Technical Commission for Aeronautics: "Air Traffic Control," Special Committee 31, Paper 27-48/DO-12, May 12, 1948, p. 24.
74. Air Coordinating Committee Air Traffic Control and Navigation Panel, Operational Policy Group: "Air Traffic Control and the National Security," December, 1950, p. 25.
75. Radio Aids to Navigation: Definition of Terms, *Proc. IRE*, **37**, 1364–1371, December, 1949.
76. Diamond, H., and F. W. Dunmore: Radio Beacon and Receiving System for Blind Landing of Aircraft, *Proc. IRE*, **19**, 585, April, 1951.
77. Jackson, W. E.: Status of Instrument Landing Systems, *CAA Tech. Development Rept.* 1, October, 1937.
78. Jackson, W. E., A. Alford, P. F. Byrne, and H. B. Fischer: Development of the Civil Aeronautics Authority Instrument Landing System at Indianapolis, *Elec. Commun.*, **18**, 285–302, April, 1940.
79. Buttner, H. H., and A. G. Kandoian: Development of Aircraft Instrument Landing Systems, *Elec. Commun.*, **22**, 179–192, April, 1945.
80. Provisional International Civil Aviation Organization: Aeronautical Telecommunications and Radio Aids to Air Navigation, Recommendations for Standards, Practices, and Procedures, P. Com 13, May, 1946, Montreal, Canada.
81. International Civil Aviation Organization: Standards and Recommended Practices, Aeronautical Telecommunications, Annex 10 to the Convention on International Aviation, June, 1949, p. 24, Montreal, Canada.
82. Watts, C. B., Jr.: Some Considerations of Wide Aperture Localizer Antennas, *CAA Tech. Development Rept.* 155, January, 1952.
83. Pickles, S.: Army Air Force Portable Instrument Landing System, *Elec. Commun.*, **22** (4), 262–294, 1945.
84. Watts, C. B., Jr.: Theoretical Consideration of an Improved Glide Path Antenna System, *CAA Tech. Development Rept.* 81, March, 1949.
85. Zero Reader Offers Way for Simplifying Cockpit, *Am. Aviation*, Oct. 15, 1948, p. 33.

86. Anast, J. L.: Automatic Flight Control, *Aero. Eng. Rev.*, **11**, 26, May, 1952.
87. Provisional International Civil Aviation Organization: "Final Report—First Session," Special Radio Technical Division, January, 1947, p. 33, Montreal, Canada.
88. Radio Technical Commission for Aeronautics: Air Traffic Control Paper 27-48/DO-12, Special Committee 31, p. 22, May 12, 1948.
89. International Civil Aviation Organization: Standards and Recommended Practices, Aeronautical Telecommunications, Annex 10, p. 18, June, 1949, Montreal, Canada.
90. Draper, C. S.: The Sonic Altimeter for Aircraft, *NACA Tech. Note* 611, p. 4, August, 1937.
91. Sandretto, P. C.: Absolute Altimeters, *Proc. IRE*, **32**, 167–173, March, 1944.
92. Radio Technical Commission for Aeronautics: "Recommended United States Policy Air Navigation-Communication-Traffic Control," Aug. 28, 1946, p. 6.
93. Special Radio Technical Division, Provisional International Civil Aviation Organization: "Final Report—First Session," January, 1947, p. 45, Montreal, Canada.
94. Warskow, Martin A.: "Airport Surface Navigation and Traffic Control," RTCA, spring meeting, Mar. 27, 28, 1951, p. 2.
95. Special Radio Technical Division, Provisional International Civil Aviation Organization: "Final Report—First Session," January, 1947, p. 45, Montreal, Canada.
96. "London Airport Ground Traffic Movement," p. 24, H.M. Stationery Office, London, 1949.
97. How Airport Surface Detection Operates, *Am. Aviation*, Mar. 31, 1952, p. 14.

Bibliography

Air-borne Direction Finders and Radiophares:

Bond, D. A.: "Radio Direction Finders," McGraw-Hill, 1944.

Busignies, H.: Automatic Radio Compass and Its Application to Aerial Navigation, *Elec. Commun.*, October, 1936.

———: Mountain Effects and the Use of Radio Compasses and Radio Beacons for Piloting Aircraft, *Elec. Commun.*, **19** (3), 1941.

Everitt, W. L., and G. E. Anner: "Communication Engineering," 3d ed., McGraw-Hill, 1956.

Framme, R. J.: "Radio Compasses, Principles, and Testing," IRE 14th Annual Convention, Sept. 22, 1939.

"Ionosphere Radio Propagation," *Nat. Bur. Standards Circ.* 462, June 25, 1948.

Massachusetts Institute of Technology: "Applied Electronics," Wiley, 1943.

South African Council for Scientific and Industrial Research: Performance of Automatic Radio Compass in High Atmospheric Noise Levels, *Research Rept.* ETR-12, 1951.

Williams, C.: Low-frequency Radio-wave Propagation by the Ionosphere, with Particular Reference to Long-distance Navigation, *Proc. IEE*, Part 3, March, 1951.

Four-course Low-frequency Radio Range and Markers:

Everitt, W. L., and G. E. Anner: "Communication Engineering," 3d ed., McGraw-Hill, New York, 1956.

Hucke, H. M.: Precipitation Static, Interference on Aircraft and at Ground Stations, *Proc. IRE*, May, 1939.

Jackson, W. E., and D. M. Stuart: Simultaneous Radio Range and Telephone Transmission, *Proc. IRE*, March, 1937.

——— and H. I. Metz: Development, Adjustment and Application of the Z Marker, *CAA Tech. Development Rept.* 14, July, 1938.

Keen, R.: "Wireless Direction Finding," 2d ed., Iliffe, 1938.

Stuart, D. M.: Circuit Design for Low-frequency Radio Ranges, *CAA Tech. Development Rept.* 23, November, 1939.

Consol:

Brown, A. H.: Consol Navigation System, *J. IEE*, Part 3A, **94** (15), 1947.

"Consol Navigation System," Third Commonwealth and Empire Conference on Radio for Civil Aviation, Conference Paper 15A, summer, 1945.

Jessel, A. H.: Range and Accuracy of Consol, *J. Inst. Nav.*, **1** (3), July, 1948.

High-frequency Direction Finding from Ground Stations:

Barfield, R. H.: Some Principles Underlying the Design of Spaced-aerial Direction Finders, *J. IEE*, **76**, 1935.

Keen, R.: "Wireless Direction Finding," 3d ed., Iliffe, 1938.

Terman, F. E.: "Radio Engineering," McGraw-Hill, 1932.

Loran:

Freas, R. R.: Airborne Loran Receiver—AN/APN-70, *Trans. IRE*, PGANE, March, 1954.

Hall, J. S.: "Radar Aids to Navigation," M.I.T. Radiation Laboratory Series, Vol. 2, McGraw-Hill, 1947.

Loran Receiver Indicator, *Electronics*, December, 1945.

Pierce, J. A.: Navigational Uses of Low Radio Frequencies, *J. Inst. Nav.*, **3**, December, 1952.

———, A. A. McKenzie, and R. H. Woodward: "Loran," M.I.T. Radiation Laboratory Series, Vol. 4, McGraw-Hill, 1948.

Smith, R. A.: "Radio Aids to Navigation," Macmillan, 1948.

Electronic Pilotage Equipment:

Allison Navigational Radar, *Aero Digest*, October, 1950.
Bell Telephone Laboratories Staff: "Radar Systems and Components," Van Nostrand, 1949.
Fink, D. G.: "Radar Engineering," McGraw-Hill, 1947.
Hall, J. S.: "Radar Aids to Navigation," M.I.T. Radiation Laboratory Series, Vol. 2, McGraw-Hill, 1947.
Ridenour, L. N.: "Radar System Engineering," M.I.T. Radiation Laboratory Series, Vol. 1, McGraw-Hill, 1947.

Electronic Aids to Dead Reckoning:

"AVQ-9 Radar Altimeter," Instruction Book, Radio Corporation of America, Camden, N.J.
Goldman, A.: Pulse-type Radio Altimeter, *Electronics*, June, 1946.
Greenwood, I. A.: Drift Determination with Radar, Using a Pulse-Doppler-Beat Method, *J. Inst. Nav.*, **6**, June, 1947.
Klass, Philip J.: Inertial Navigation, special report in *Aviation Week*, January 2, 9, 16, and 23, 1956.
Langdon, H. F.: Navigating by Altimeter, *Air Transport*, May, July, 1946.

Very-high-frequency Omnidirectional Range (VOR):

Anderson, S. R., and H. F. Keary: Polarization Errors of Two Different Omni-range Antenna Arrays, *CAA Tech. Development Rept.* 63, March, 1950.
Gross, F. J., and M. Kincaid: Procedure for Calibrating Collins 51R-1 Navigation and BC733D Localizer Receivers, *CAA Tech. Development Rept.* 122, September, 1950.
Hurley, H. C., S. R. Anderson, and H. F. Keary: CAA VHF Omnirange, *CAA Tech. Development Rept.* 113, June, 1950.
Rambo, W. R., J. S. Prichard, D. P. Duffy, and R. C. Wheeler: "Summary Report on Evaluation of Omni-bearing-distance System of Air Navigation," Airborne Instruments Laboratory, Report 540-1, October, 1950.

Distance-measuring Equipment:

Borden, R. C., C. C. Trout, and E. C. Williams: UHF Distance Measuring Equipment for Air Navigation, *CAA Tech. Development Rept.* 114, June, 1950.
Burgmann, V. D.: Distance Measuring Equipment for Aircraft Navigation, *J. IEE*, **96**, 1949.
Busignies, H.: High-stability Radio Distance-measuring Equipment for Aerial Navigation, *Elec. Commun.*, **25**, September, 1948.
"DME System Characteristics (Transition Period)," Report of RTCA SC-40, Dec. 15, 1948.
Sandretto, P. C.: Rho-Theta System of Air Navigation, *Elec. Commun.*, **27**, December, 1950.

Some Avigational Aids for the Short-distance En-route Zone:

Adams, P. R., S. H. Dodington, and J. H. Herbst: "First Tests on Navar System for Aerial Navigation and Air Traffic Control," IRE Convention, 1947.
Cleaver, R. F.: Development of Single-receiver Automatic Adcock Direction-finders for Use in the Frequency Band 100–150 Megacycles per Second, *Elec. Commun.*, **25**, December, 1948.
"DME System Characteristics (Transition Period)," Report of RTCA SC-40, 121-48/DO-24, Dec. 15, 1948.
Hopkins, H. G., and F. Horner: Direction Finding Site Errors at Very High Frequencies, *J. IEE*, Part 3, **96**, 1949.
Lee, J. M., R. G. Palmer, and B. M. Lohr: Investigation to Determine the Characteristics of Horizontal and Vertical Polarization for Very-high-frequency Two-course Visual Radio Ranges, *CAA Tech. Development Rept.* 58, July, 1947.
Lyman, J., and G. Litchford: Distance Measuring Equipment for the Terminal Area, *Aero. Eng. Rev.*, **9** (5), May, 1951.
——— and ———: Precision Omnidirectional Radio Range for the Terminal Area, *Aero. Eng. Rev.*, **9** (6), June, 1950.
Smith, R. A.: "Radio Aids to Navigation," Cambridge, 1948.
Tuller, W. G.: Theoretical Limitations on the Rate of Transmission of Information, *Proc. IRE*, **37**, May, 1949.

Tacan:

Elec. Commun., **33** (1), March, 1956 (Technical Journal of the International Telephone and Telegraph Corporation).

Airport Surveillance Radar:

Anderson, L. J., J. P. Day, C. H. Freres, and A. P. D. Stokes: Attenuation of 1.25-centimeter Radiation through Rain, *Proc. IRE*, **35**, April, 1947.
Garner, W. R., and F. Hamburger, Jr.: Detectability and Discriminability of Targets on a Remote Projection Plan-position Indicator, *Proc. IRE*, **35**, November, 1947.
Haeff, A. V.: Minimum Detectable Radar Signal and Its Dependence upon Parameters of Radar Systems, *Proc. IRE*, **34**, November, 1946.
Norton, K. A., and A. C. Omberg: Maximum Range of a Radar Set, *Proc. IRE*, **35**, January, 1947.
Payne-Scott, Ruby: Visibility of Small Echoes on Radar P.P.I. Displays, *Proc. IRE*, **36**, February, 1948.

Randall, A. W., and J. S. Marshall: Bright Display Equipment for Surveillance Radar, *CAA Tech. Development Rept.* 173, June, 1952.

———, C. C. Trout, J. E. Herrmann, and P. H. Leake: Performance of S-band and L-band Surveillance Radars during Rain and Cloud Conditions, *CAA Tech. Development Rept.* 92, April, 1949.

Ridenour, L. N.: "Radar System Engineering," M.I.T. Radiation Laboratory Series, Vol. 1, McGraw-Hill, 1947.

Robertson, S. D., and A. P. King: Effect of Rain upon the Propagation of Waves in the 1- and 3-centimeter Region, *Proc. IRE*, **34**, April, 1946.

Soller, T., M. A. Starr, and G. E. Valley, Jr.: "Cathode-ray Tube Displays," M.I.T. Radiation Laboratory Series, Vol. 22, McGraw-Hill, 1948.

Fixed-beam Low-approach Systems:

Anderson, W. G., and E. H. Fritze: Instrument Approach System Steering Computer, *Proc. IRE*, **41**, February, 1953.

Caporale, P.: CAA Instrument Landing System, Part I, *Electronics*, **18**, February, 1945.

Davis, R. C.: Glide Slope Receiver, *Elec. Commun.*, **29**, September, 1952.

Flush Mounted Antennas for Military Aircraft, *Tele-Tech*, October, 1952.

Hampshire, R. A., and B. V. Thompson: ILS-2 Instrument Landing Equipment, *Elec. Commun.*, **27**, June, 1950.

Metz, H. I.: CAA-RTCA Instrument Landing System, *CAA Tech. Development Rept.* 35, October, 1943.

Overbury, F. G.: Airborne Receivers and Test Gear for Instrument Landing Systems, *Elec. Commun.*, **29**, June, 1952.

Radar Low-approach System:

Engel, J. S.: Landing Aircraft with Ground Radar, *Elec. Commun.*, **24**, March, 1947.

Hall, J. S.: "Radar Aids to Navigation," M.I.T. Radiation Laboratory Series, Vol. 2, McGraw-Hill, 1947.

Watson, C. W.: Ground-controlled Approach for Aircraft, *Electronics*, November, 1945.

Landing Altimetry:

Capelli, M.: Radio Altimeters, *S.R.E.C. Record*, No. 5, October, 1953.

Espenschied, Lloyd, and R. C. Newhouse: A Terrain Clearance Indicator, *Bell System Tech. J.*, **18**, 1939.

Pasek, D. M., and W. J. Shanahan: "Flare Out Unit AN/APN-71, An Aid to Aircraft Instrument Landing," Abstract of Papers delivered at the 1951 Conference on Airborne Electronics, IRE, May, 1951.

Sharpe, B. A.: Low-reading Absolute Altimeter, *J. IEE, Radiocommunication*, **94**, Part IIIA, 1947.

Airport-surface-detection Radar:

Anderson, Clair M., and Marvin H. Yost: Evaluation of Airport Surface Detection Equipment, Model AN/MPN-7 (XW-1), Part II, *CAA Tech. Development Rept.* 176, November, 1952.

Blount, Edward M., Samuel L. Kodes, Hugh A. Kay, and Raymond E. McCormick: Evaluation of Airport Surface Detection Equipment, Model AN/MPN-7 (XW-1), Part I, *CAA Tech. Development Rept.* 175, November, 1952.

CHAPTER 27

WIRE TELEGRAPHY

By I. S. Coggeshall[1]

The inclusion of chapters on wire communication in a handbook on radio engineering is indicative of the elimination of former boundaries separating wire and wireless, of the identity of much present technique and apparatus, and of the prevalence of end-to-end operation of radio and wire links in "through" telephone and telegraph circuits and services. Radar and the video arts have established radio interest in telegraphic signal elements, pulsing techniques, and codes, and in the extension of their application to industrial automation. The problems of nonlanguage nonredundant information and business data handling and processing serve further to integrate wire and radio telegraphy and telephony.

The fact that the two "wire" chapters (Chaps. 27 and 28) are the last to be added to this handbook accounts for the omission here of expositions commonly appearing in wire-communication texts, or of material already appearing elsewhere in this book, particularly in Chap. 1, the basis of all electrical communication; Chaps. 2 to 5, on circuits of R, L, and C; Chaps. 13 and 14, on electrical measurements; Chap. 17, power-supply systems; Chap. 23, facsimile; and Chap. 24, code transmission and reception in radio. Portions of these, and the other chapters on audio, radio, and electronic circuitry, are basic to modern wire telegraphy and telephony and in certain instances were derived from wire practices and discoveries. Repetition of this included material has been avoided in the wire chapters.

Since engineers whose major concern is specialization in wire telegraphy per se, apart from radio, may already consult texts catering to their purposes,[2] such treatments are not matched here in scope or detail. Omitted also are historical, obsolete, and obsolescent material except where a knowledge of it seems pertinent to new applications or to trends the industry is taking.

Demarcation of the scopes of the two wire chapters, respectively, has necessarily been somewhat arbitrary in view of the fact that the techniques of "record" and "voice" transmission have tended to become unified by their simultaneous occupancy of wire facilities. At one time it would have been possible to say that the use of codes by telegraphy set it apart from telephony, but today speech and music may be encodified during transmission. Telegraphic techniques were once exclusively "narrow-band" compared with telephony; today both appear often as rather similar modulations of megacycle-carrier frequencies. Among the "weak-current" techniques, telegraphy used to employ heavier currents and higher voltages (chiefly d.c.) than telephony, but these are no longer characteristic.

The dividing line adopted, therefore, has been directed to the convenience of the reader. In general, telephony has been made to embrace the *voice band*, 200 to 3,500 cps, and everything above it; telegraphy, the *subvoice* frequencies and the subdividing of the voice band for code. When telegraphic and telephonic modulations, as they often do, ride the same carriers at the same time, the treatment in this handbook will be found under telephony.

[1] The Western Union Telegraph Company, New York.

[2] The principal available texts appear as references at the end of this chapter.

TRANSMISSION THEORY

1. Wire-transmission Objectives. The objective of electrical transmission by wire is the same as that of radio propagation: to create at a distance from a transmitter a disturbance in an electromagnetic field which can be detected by a receiver in such a way that intelligence shall have been conveyed. In both the wire and radio modes of propagation a small portion (milliwatts or microwatts) of the total electrical power sent into the medium of transmission suffices, after local amplification at point of reception or repetition, to actuate the receiver or repeater.

The distinguishing characteristic of wire transmission is its employment of a conductor-dielectric system of continuous material boundaries to guide the energy of the electromagnetic waves rather than to launch it into space. This makes it possible to dispense with r-f waves as carriers of audio and subaudio intelligence modulations, since wire systems transmit energy at all frequencies from their practical upper cutoff frequencies down to zero cps (d.c.).[1]

2. Applicability of Maxwell's Equations. Maxwell's equations (Chap. 1, Sec. 131) are basic to mathematical interpretations of electromagnetic fields in five categories, with which communication theory and engineering are consistent. The first, the *static* case, is unimportant to transmission. The second, the *steady-state* case for stationary (d-c) field, applies, after the passage of an initial wave-front transient, to transmission lines so short, or to frequencies of polarity reversal so low, that the time duration of the shortest impulses employed exceeds the total time of propagation of the transient.

Local-circuit-network theory, based upon lumped linear parameters, typifies Maxwell's (third) *quasi-steady-state* case for closed circuits. So do the "telegraphist's equations" (Sec. 8), applicable to intercity wire transmission lines with distributed parameters, operated at all but the very high frequencies. Thus ordinary a-c network theory may be applied to line problems involving transients, as met with in transmission of telephone carrier over cable pairs, where, because of the higher audio frequencies and relatively slow wave-propagation velocities, sections between repeaters are long enough to span a succession of signal impulses simultaneously, and in transmission over submarine telegraph cables, where despite the low frequencies the same multiple-wave succession exists because of extreme section lengths.[2]

As frequencies are raised[3] past the point (1) where displacement currents are no longer negligible in comparison with conduction and convection currents; (2) where the flow of electrons is no longer streamlined along the conductors but has developed a perpendicular, radiating component;[4] or (3) where, because the dimensions of the circuit exceed about a wavelength of oscillation, the instantaneous electromagnetic field is not uniform throughout—in these events the solutions[5] of Maxwell's equations pass through the (fourth) *quasi-steady* case for open circuits[6] like wave guides[7] into the (fifth) *general* case for dynamic phenomena which includes r-f radiation into space.

Wire transmission and radio propagation are thus seen to be governed by unified physical concepts, with shifting emphasis as the spectrum of frequencies and the circuit electrical parameters and geographical configurations come into play.

[1] Weber, E., Traveling Waves on Transmission Lines, *Elec. Eng.*, **61**, 302, 1942.

[2] Coggeshall, I. S., Testing and Regulating—Overseas, *Western Union Tech. Rev.* (abbreviated hereafter: *WUTR*), **5**, 23, 1951.

[3] Race, H. H., and C. V. Larrick, High-frequency Coaxial-line Calculations, *Trans. AIEE*, **61**, 526, 1942.

[4] Laport, E. A., Open-wire Radio-frequency Transmission Lines, *Proc. IRE*, **31**, 271, 1943.

[5] Frankel, S., Equations for Generalized Transmission Lines, *Elec. Commun.*, **23**, 329, 1946.

[6] Clavier, A. G., Reciprocity between Generalized Mutual Impedances for Closed or Open Circuits, *Elec. Commun.*, **27**, 152, 1950.

[7] Schelkunoff, S. A., Generalized Telegraphist's Equations for Waveguides, *Bell System Tech. J.*, **31**, 784, 1952. Morgan, S. P., Mathematic Theory of Laminated Transmission Lines, *Bell System Tech. J.*, **31**, 883, 1121, 1952. Vaage, E. F., Transmission Properties of Laminated Clogston Type Conductors, *Bell System Tech. J.*, **32**, 695, 1953. Miller, S. E., Waveguide as a Communication Medium (bibliog.), *Bell System Tech. J.*, **33**, 1209, 1954. Grieg, D. D., and H. F. Engelmann, Microstrip—A New Transmission Technique for the Kilomegacycle Range, *Proc. IRE*, **40**, 1644, 1651, 1952.

3. Propagation of Waves on Wires. Electromagnetic waves on wires may be thought of substantially as though radiated through space but bound to, hence guided by, the "line." A transmission line consists of three essential elements: two conductors of some sort and insulation separating them. Among typical boundary configurations of these three elements may be mentioned: (1) two identical parallel wires of copper, iron, or copper-clad steel, on a pole line, the dielectric being air and glass insulators; (2) a pair of small-gage copper wires in a paper-insulated underground cable; (3) a configuration consisting of one aerial wire on a pole line, the ground or earth constituting the second conductor, and between them air dielectric, wooden poles, and glass insulators; (4) a submarine cable in which a gutta-percha or polyethylene dielectric surrounds one conductor of stranded copper and separates it from a second conductor consisting of steel-armor wires in shunt with sea water; (5) other types of coaxial systems consisting typically of a copper wire within a copper tube, the dielectric being ceramic or plastic supports in air or gas.

Consonant with the two-conductor-and-dielectric concept, applying alike to all these typical line configurations, wire transmission employs, as radio propagation through space also does, familiar features of the kinetic-potential, energy-exchanging, wave-producing mechanism of classical physics. Any increase or decrease in electric-field intensity, propagated outward from a source of disturbance in the circuit, produces a wave front giving rise to variations of magnetic-field intensity; these in turn give rise to further variations of the electric field, and so on. Once established, the original disturbance is sustained by the mutual interaction of electric and magnetic components of the field as the wave travels onward. Free electrons in the conductors are swept into motion (which is another way of saying that the electric charges from the emf source are spread) by the energy exchange taking place as the wave front progresses.

If it is bipolar battery which is applied, it results in the spreading of oppositely poled charges along the respective conductors. The spreading of these opposite positive and negative charges to the far ends of the conductors marks a wave front of instantaneous transient disturbance and initial electron movement at any point, in the directions determined by the polarized fields.

Conduction and convection currents, coinciding with net electron flow in the conductors, are coupled inductively and capacitatively across the dielectric between them to determine circuit constants of propagation velocity, power attenuation, frequency discrimination, and so on, in accordance with circuit theory.

4. Convection of Electrons in Wires. The wave front of electromagnetic-field disturbance, in and about conductors separated by dielectric, is the locus of initiation of the convection of electrons, constituting electric current in the wires. The disturbance may occur at and be propagated almost instantaneously throughout the length of the line from any point on the circuit but in practice usually is associated with a change of polarity or value (including zero, or ground, potential) of applied emf at some point, or with the making or breaking of a charged circuit at any point.

The establishment and maintenance of electron flow is fundamental to the transmission of communication power. The electron stream from cathode to plate in a diode, at velocities of the order of a tenth or more of the speed of light, may properly be thought of as sustained by corresponding flow of the same number of electrons per unit of time through the B battery and the completing circuit of copper wire. But within the copper wire itself, free electrons, which have a net haphazard velocity of about 35 miles per sec at ordinary temperatures, undergo so many atomic collisions that their progressive convection around the circuit under the influence of applied emf is only a fraction of a centimeter per second.[1]

This slow rate of electron progression in a closed circuit might initially seem anomalous with the "lightning" velocities of propagation of wave fronts over transmission lines, varying from, say, 10,000 miles per sec for loaded cable pairs to 180,000 miles per

[1] Morecroft, J. H., "Principles of Radio Communication," 3d ed., p. 12, Wiley, 1933. Darrow, K. K., Electronic Conductors, *Bell Lab. Record*, **33**, 85, 1955.

sec for open wire lines. However, by analogy it will be seen that two consistent independent velocities are involved. If two identical long, taut, parallel horizontal chains were anchored at one end to opposite arms of a lever, and by movement of the lever one chain were suddenly *released* a little and the other simultaneously *pulled* a little more taut, perceptible waves would travel in the same direction along both chains at identical velocities which would be a function of the characteristics of stretched chain and independent of the velocity or direction of the communicated influences, whether they happen to be "pull" or "release."

By analogy and as an electrical transmission-line illustration the wavelength of 1,000-cps a.c. on a 200-mile open-wire pair is 177 miles—indicative that (1) electrons are flowing in opposite directions in the same conductor 88.5 miles apart; (2) in the case of balanced conductors, at any point on the line, including both terminations, the flow of electrons in one conductor is opposite in direction to the flow of electrons in the other; yet (3) the wave-front progression is in a single direction, outward from the point where the source emf is applied, at a velocity of 177,000 miles per sec. It follows that, at the place and instant of application or variation of emf, and in the very process of launching the wave front forward at nearly the speed of light, the electrons in one of the conductors are progressing rather slowly backward into the positive pole of the emf source, somewhat as in the stretched-chain analogy.

The electron-flow concept is also useful in demonstrating the palpable flow of current (at all times and places in a direction opposite to the flow of electrons) once an electron drift has been initiated by the rapid advance of the disturbance front in the electric and magnetic fields along the conductors and the continuation of that drift for so long as the polarity and value of the applied emf remain constant to maintain it.

During such a d-c interval, before a second transient occurs to disturb the fields, there is no further movement in them. In the absence of moving-field interactions there can be no inductive or capacitative effects. Circuit parameters are reduced to series resistance in the conductors and shunt leakage conductance between them. The electron progressions set up by the wave front continue, sustained by the steadily applied emf and in the directions predisposed by the passage of the wave front: forward, out of the battery, into one conductor; backward, into the battery, from the other; and in amounts equal in the two conductors at the emf source. The transmission line, during this interval, obeys Ohm's law for d.c., as Maxwell's special (second) *steady-state* case for stationary fields; it is a circuit for the current but passive as a transmission line, awaiting arrival of a new wave front to alter the direction or volume of drift of electrons throughout both conductors.

5. Uniform Transmission Lines. A line of any constant conformation throughout its length may be thought of as composed of incremental lengths having identical electrical properties. Any convenient unit of length may be chosen, but on long "ground-return" lines it is usual to describe and tabulate the characteristics at a given frequency in terms of their values per mile of geographical distance between the ends and, on metallic-loop lines, in terms of values per loop-mile. The lumped circuit elements are: series resistance R of both conductors, in d-c ohms; inductance L of the conductors in henrys; capacitance C between the conductors, in microfarads; and a leakage resistance, conductance G in micromhos, between the two conductors, as shown in Fig. 1. Where one of the conductors is the ground it is assumed to have a resistance of zero ohms per mile.

Such a line, having uniform characteristics throughout its length, is said to be a "smooth" line, having no impedance discontinuities which would give rise to reflections and consequent power losses. The uniform configuration of the incremental sections gives each one its "characteristic" impedance; matched impedances ensure maximum power transfer from section to section; and if the line is terminated at both ends in matching characteristic impedances, the full transmission value of the line, at any given frequency, is realized.

Although conductors in pairs are usually matched as to material and size, a line otherwise composed (for example, of one copper and one iron conductor or of coaxial

conductors or of one wire and ground) is still a smooth or uniform line if it contains no discontinuities between incremental sections.

Perfectly uniform lines are seldom encountered but in most cases lines are good enough to justify the use of calculations based on formula. Extreme requirements for uniformity sometimes have to be met in engineering—*e.g.*, in ocean-telephone-cable manufacture.[1]

6. Application of Electromotive Force. Signaling emf impressed upon communication wires is ordinarily either applied directly to the two conductors (of whatever conformation, including one wire and ground) or is used to *modulate* (that is, vary the envelope form of) an a-c source in amplitude or frequency. For direct application, the most common form of wave shape is *flat-top* d.c., either interrupted by intervals of zero emf or undergoing polarity reversals, in either case generated commonly by electromagnetically or cam-operated relay contacts, tuning-fork contacts, the sweep of brushes across segmented commutators or faceplates, by hand-operated keys or pole changers, or by circuits containing unidirectional circuit elements such as rectifiers or thyratrons. The flat-top form in certain cases is modified by circuit elements, sometimes for a special purpose such as decreasing the higher-frequency harmonic content to prevent extraneous interference. Sine-wave shapes have upon occasion been used for signaling emf, modulated as to amplitude levels (and hence constituting a carrier system where the carrier and modulating frequencies were identical);[2] but in general the principal use of the sinusoidal wave form of emf is for the to-be-modulated carrier, and the ordinary generator is the tube or transistor oscillator, with or without piezoelectric frequency control.

Any desired wave form may be generated for telegraphic use. In the case of certain submarine cables the sudden discharge of a capacitor into the line produces a peaked sending pulse. Such peaked-pulse forms may be position-modulated for telegraphic as well as telephonic transmission.

In telephony the wave form is generated by electrical coupling with acoustical variations produced by voice or musical instruments. It is treated in accordance with its content of Fourier-series components of sine-wave form.

7. Transient vs. Quasi-steady-state Solutions. When any form of emf is suddenly applied to or removed in whole or in part from a circuit, the upset of the equilibrium of the circuit gives rise to a *transient* the characteristics of which may be calculated by the use of nonperiodic functions, expressed most rigorously in terms of the *Fourier integral*.[3] However, a sufficiently accurate and more convenient engineering approach is commonly made, based upon dynamic equilibrium of a circuit subject to applied a-c emf over a range of frequencies representing the *Fourier-series analysis* of the wave shape important to the transmission; communication-transmission theory is practically founded on such quasi-steady-state relationships.[4]

A complete expression of the effect upon a receiver of emf impressed at the sending end not only has to take into account the transmission losses contributed by the line itself but must consist of the product of additional terms showing the modifications introduced by the sending and receiving apparatus and by reflections from both ends and between ends.[5] Where terminations can be made in the characteristic impedance of the line itself, the reflection factors approach unity. Usually the characteristics of the line are the predominant ones in the calculation of circuit transmission.

8. Applicability of Telegraphist's Equations. For conventional dissipative wire transmission lines including coaxial cables, Maxwell's equations can be converted into accurate forms of "telegraphist's equations" (now more often called *transmission-line*

[1] Kelly, M. J., G. Radley, G. W. Gilman, and R. J. Halsey, A Transatlantic Telephone Cable, *Trans. AIEE*, **74**, Part 1, 124, 1955.

[2] *Cf.* Squier-Bruce code (Sec. 28).

[3] Campbell, G. A., and R. M. Foster, "Fourier Integrals for Practical Applications," Van Nostrand, 1948.

[4] Coughlin, R. F., Fourier Series Analysis of Local Circuit Telegraph Waves, *WUTR*, **9**, 33, 1955.

[5] Osborne, H. S., "The Principles of Electric Circuits Applied to Communication," Bell System Monograph, 1928. Milnor, J. W., Submarine Cable Telegraphy, *Trans. AIEE*, **41**, 20, 1922.

equations) for transmission sections of the form in Fig. 1. It has recently been shown[1] that the equations involving lumped constants of R, L, C, and G are approximately consistent with Maxwell's field theory at much higher frequencies than indicated by their derivation from quasi-steady-state fields and may be generalized in a manner

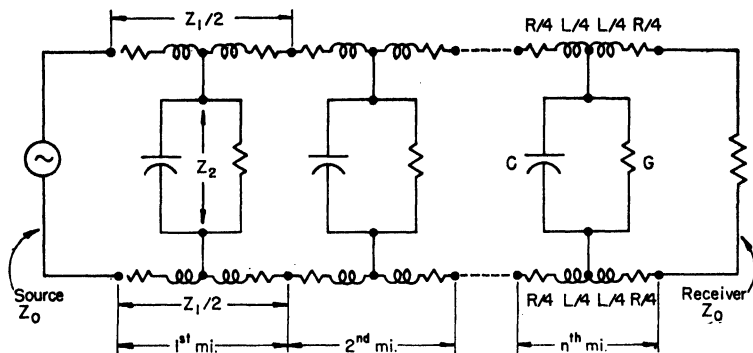



FIG. 1. Uniform transmission-line pair—1st, 2d, nth miles. For given frequency, lumped circuit constants: series parameters, R, L, $\mathbf{Z}_1$ per loop-mi.; shunt parameters, C, G, $\mathbf{Z}_2$ per mi.; characteristic and matched source and receiver impedances $\mathbf{Z}_0$.

suitable for solving boundary-value problems, thus bringing together theories which have coexisted for many years.

The classical form of the telegraphist's equations for two parallel conductors, each carrying instantaneous current $I_i(z)$ at any point z on the line and separated by a transverse instantaneous voltage of $V_i(z)$ (Fig. 2), is

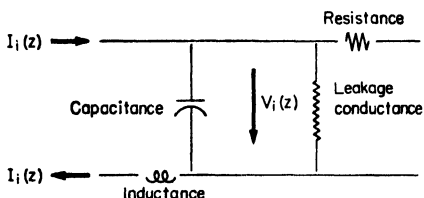



FIG. 2. Derivation of location (z) functions of instantaneous V_i and I_i values along parallel conductor transmission line having distributed R, L, C, and G computed under static conditions.

$$\frac{dV_i}{dz} = -RI_i - L\frac{dI_i}{dt}$$

and

$$\frac{dI_i}{dz} = -GV_i - C\frac{dV_i}{dt}$$

where R, L, G, C are the distributed values per unit length along the line, computed from static considerations. If sinusoidal I and V are expressed as exponentials, *i.e.*, $I(t) = Ie^{j\omega t}$, $V(t) = Ee^{j\omega t}$, the z functions become

$$\frac{dV}{dz} = -I(R + j\omega L) \qquad \text{and} \qquad \frac{dI}{dz} = -V(G + j\omega C)$$

$$V(z) = ke^{Pz} \qquad \frac{d^2V}{dz^2} = ke^{Pz}P^2$$

$$V(z) = k_1e^{-Pz} + k_2e^{Pz}$$

$$I(z) = \frac{1}{Z_0}(k_1e^{-Pz} - k_2e^{Pz})$$

where k_1 and k_2 are arbitrary constants of second-degree differential equations in V or I.

[1] Schelkunoff, S. A., Conversion of Maxwell's Equations into Generalized Telegraphist's Equations (bibliog.), *Bell System Tech. J.*, **34**, 995, 1955.

From these equations the *propagation constant*[1] (the factor reflecting logarithmic decay of current over each unit length of line) is found to be

$$P = \sqrt{(R + j\omega L)(G + j\omega C)} = A + jB$$

[*cf.* Chap. 5, Eq. (111)], A being the *attenuation constant* (indicating the loss of current magnitude) and B the *wavelength constant* (indicating the increase of phase shift due to time required for energy to traverse unit length of line).

Also, the *characteristic impedance* (sometimes called *iterative* or *surge* impedance) is

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}}$$

Note that the propagation constant and the characteristic impedance disclose product and quotient relationships of the parallel and branched shunt impedances of Fig. 1.

9. Transmission-line Calculations. Tables of transmission-line values are available, recording the results of measurements and calculations based upon standard line-pair configurations at various frequencies from zero to practical upper limits, and for aerial wire line of various pole-pin spacings and for inductively loaded[2] and nonloaded cable. The resistance values go up with frequency, because of *skin effect;* they go down with an increase of conductor cross-sectional size at all frequencies; they rise with ambient temperature and vary with the specific resistance of the conductor material, usually copper, iron, or copper-clad steel. Inductance values go down as frequencies go up and as wire size is increased, and are lower for 8-in. separation on aerial lines than for 12-in. and for the 18-in. pole-pair positions on the crossarms. Capacitance values depend upon contiguity of an aerial pair to other wires on the line and to ground; tables usually specify values for 40-wire lines. Capacitance values are independent of frequency but go up with an increase in wire size and if the line becomes wet. Leakance values, in micromhos per loop-mile, are subject to radical increases with frequency rise and especially with wetting of aerial line, so that the tables contain dry and wet values.

Similar tables for paper-insulated exchange, toll, and spiral-four toll-entrance cable disclose entirely different pair values,[3] as might be assumed from the different structure for aerial wire, but follow similar patterns; effective resistance is affected by proximity of conductors; inductance is low; reactance to temperature variations may be serious and is combated by use of underground cable. Leakage conductance is unaffected by ambient moisture and is further combated by the use of dry gas above atmospheric pressure. Lumped inductive loading of various tabulated types enters the picture to increase the resistance and inductance per loop-mile and produces an upper cutoff frequency not experienced in the case of open-wire or nonloaded cable pairs.

Besides the lumped-circuit constants, the available tables give 1,000-cps values for other calculated and measured line characteristics for aerial wire and toll cable.

An example will show how the values are calculated; and typical cases are shown in Table 1.

1. For 1,000 cps, $\omega = 2\pi f = 6{,}283$.
2. Given R, L, C, and G, $(R + j\omega L)$ and $(G + j\omega C)$ may be calculated, each expressed in polar coordinates in the form of a vector magnitude and an angle.
3. The *characteristic impedance* Z_0 is the square root of the quotient of these respective values, also expressed as a vector and a negative angle. This is convertible to the rectangular coordinates of R and X, in ohms.

[1] Clarke, J. L., Some Novel Expressions for the Propagation Constant of a Uniform Line, *Bell System Tech. J.*, **25**, 156, 1946.

[2] Shaw, T., The Evolution of Inductive Loading for Bell Telephone Facilities, *Bell System Tech. J.*, **30**, 149, 447, 721, 1221, 1951.

[3] Wilson, Leon T., "Factors in the Transmission Characteristics of Cable Circuits," Coles Signal Laboratory, Western Union Committee on Technical Publication, New York, 1952.

Table 1. Calculated Constants for Transmission Lines of Typical Configurations

	Unit	Loaded paper cable pair[a]	Nonloaded aerial pair[b]	WU type G open-wire carrier[b]	Submarine cable, single cond'r[c]
Length of section	Miles[d]	50	10	150	2,000[d]
Wire gage and diam	AWG, in.	No. 19/0.036	No. 8/0.128	No. 9/0.114	—/0.188
Distributed constants per loop-mile:					
Resistance R	Ohms	89.7	6.74	18.5	1.695
Inductance L	Henrys	40.0×10^{-3}	3.53×10^{-3}	3.49×10^{-3}	3.85×10^{-3}
Capacitance C	Farads	0.062×10^{-6}	0.0087×10^{-6}	0.0084×10^{-6}	0.380×10^{-6}
Conductance G	Mhos	1.5×10^{-6}	0.29×10^{-6}	8.4×10^{-6}	0.0022×10^{-6}
Frequency f	Cps	1,000	1,000	30,000	10[e]
$\omega = 2\pi f$	....	6,283	6,283	188,500	62.8
Characteristic impedance Z_0	$\angle\theta$	$827.5\angle-9.72$ deg	$650\angle-8.32$ deg	$646\angle-0.66$ deg	$268\angle-40.94$ deg
Z_0: resistance r	Ohms	806	643	646	202
Z_0: reactance x	Ohms	140.8	94	7.40	176
Constants per mile:					
Propagation P	$\angle\alpha$	$0.3222\angle80.67$ deg	$0.0356\angle81.39$ deg	$1.021\angle88.05$ deg	$0.00639\angle49.06$ deg
Attenuation A	Nepers	0.0556	0.00533	0.01695	0.00419
Phase shift B	Radian deg	0.3174/18.2 deg	0.0352/2.02 deg	1.021/58.5 deg	0.00483/0.28 deg
Phase shift total length	Deg	910 deg	20.2 deg	8,775 deg	553 deg
Wavelength λ	Miles[d]	19.79	178.5	6.154	1,300[d]
Wave velocity W	Miles/sec[d]	19,790	178,500	184,600	13,000[d]
Cutoff frequency f_c	Cps	5,700			
Sending:					
Rms emf E_s	Volts	2.9	2.56	2.54	50[f]
Rms current I_s	Amp	3.5×10^{-3}	3.94×10^{-3}	3.94×10^{-3}	0.187
Rms power P_s	Mw	10[f]	10[f]	10[f]	7,100
Receiving:					
Rms emf E_r	Volts	0.18	2.43	0.2	0.0114
Rms current I_r	Amp	220×10^{-6}	3.73×10^{-3}	310.2×10^{-6}	42.9×10^{-6}
Rms power P_r	Watts	39×10^{-6}	8.96×10^{-3}	62.0×10^{-6}	0.369×10^{-6}
Power ratio P_s/P_r	—/db	256 = 24 db	1.12 = 0.49 db	161 = 22.1 db	19.24×10^{6} = 72.8
Transmission equivalent	db/mile	0.48	0.0462	0.147	0.0364[d]

Paper-cable example adapted from "Principles of Electricity Applied to Telephone and Telegraph Work," courtesy of Bell System; others, courtesy of A. Boggs, Mrs. C. U. Watts, and C. H. Cramer, Western Union Telegraph Co.

[a] Copper side pair in cable, loading H-44-S; coils M-2, M-3, spaced 1.135 miles, 17.64 coils per wavelength; 4.1 ohms, 0.043 henry per section.

[b] 11.5- or 12-in.-spaced copper aerial conductors in metallic pair, normally transposed; dry.

[c] Nonloaded, armored submarine telegraph cable, simplex operation, single core; 650 lb copper, 375 lb gutta-percha per nautical mile.

[d] Submarine cable (last column only) in nautical miles.

[e] 10 cps in terms of received double bits, equivalent to 20 cps transmitted into cable.

[f] Typical values.

4. The *propagation constant* P is the square root of the product $(R + j\omega L)(G + j\omega C)$ and is expressed as a vector magnitude and a positive angle. The rectangular coordinates are the *attenuation constant* A and the *phase shift* B (in radians per mile)—respectively the cosine and the sine of the angle times the magnitude [*cf.* Chap. 5, Eq. (102)].

5. The total phase shift for the circuit is its loop mileage times the phase constant B per mile, expressed in radians and convertible to degrees.

6. The wavelength of the 1,000-cps wave on the pair is $2\pi/B$ miles.

7. The velocity W of the wave front is $f\lambda$ miles per sec.

8. The line propagation time is $B/2\pi f = 1/W$ sec per mile.

9. To investigate relationships of sent and received voltages, currents, and power losses, one assumes a sending power P_s of, say, 10 mw.[1]

10. The sending voltage E_s will be the square root of 0.01 watt times the magnitude of Z_0 vector over the cosine of its angle (the cosine will be 1 if Z_0 is a pure resistance).

11. The sending current $I_s = E_s/Z_0$.

[1] A numerical example is worked out in "Principles of Electricity Applied to Telephone and Telegraph Work," pp. 161–163, a nonpublished text of the American Telephone and Telegraph Company, 1953.

12. The received current I_r will have been reduced by the action of the shunt paths across each of n sections of line:

$$\log_e \frac{I_r}{I_s} = -nA \qquad \text{or} \qquad 2.303 \log_{10} \frac{I_s}{I_r} = nA$$

from which I_r can be calculated, the other values being known.

13. The magnitude of the received voltage E_r will have been similarly reduced and can be similarly calculated.

14. The received power $P_r = E_r I_r \cos \theta$, where θ is the same angle as in step 10. Voltage and current preserve the same relative phase relationships along the line regardless of the results from step 5.

15. The decibel loss on the line $= 10 \log_{10} (P_r/P_s)$, and the transmission equivalent, or decibel loss, per mile is that amount divided by the length of the line (loop) in miles. The loss expressed in decibels per mile is also obtainable directly by applying a factor of 8.68 to the attenuation constant A (step 4) expressed in nepers. $8.68 \text{ db} = 20 \log_{10} e \text{ db} = 1$ neper.

16. For loaded cable pairs the *cutoff frequency* f_c is approximately $1/(\pi \sqrt{LC})$, where L in henrys is the inductance of the loading coils and C in farads is the total capacitance of the length of circuit between loading coils. Useful range $\leqq 0.75f_c$. Upper practicable design limit $f_c = 30{,}000$ cps. Uniform line equations must be applied to lumped loading conditions with caution.

Table 1 gives values for some selected typical uniform telegraph lines calculated in accordance with the foregoing steps.

Table 2 expresses, for purposes of tentative design of lines of different kinds, exact and approximate noncomplex expressions for A, B, R, and X, derived from the complex equations for P and Z_0.

Table 2. Design Formulas—Transmission Lines

Basic Relationships:

$P = A + jB = \sqrt{(R + j\omega L)(G + j\omega C)}$	$Z = r + jx = \sqrt{\dfrac{R + j\omega L}{G + j\omega C}}$
Nonloaded Small-gage Cable, Up to Fairly High Frequencies	Loaded Cable; High-frequency Cable; Open-wire Line—V-f to High Frequency
Conditions: $R > \omega L \quad \omega C > G$ $\Delta_1 = \dfrac{\omega L}{R} \quad \Delta_2 = \dfrac{G}{\omega C}$	Conditions: $\omega L > R \quad \omega C > G$ $\Delta_1 = \dfrac{R}{\omega L} \quad \Delta_2 = \dfrac{G}{\omega C}$
$A = \sqrt{\dfrac{R\omega C}{2}[\sqrt{(1+\Delta_1^2)(1+\Delta_2^2)} + \Delta_2 - \Delta_1]}$	$A = \omega\sqrt{LC}\sqrt{\tfrac{1}{2}[\sqrt{(1+\Delta_1^2)(1+\Delta_2^2)} + \Delta_1\Delta_2 - 1]}$
$A \doteq \sqrt{\dfrac{R\omega C}{2}}\left(1 + \Delta_2 - \Delta_1 + \dfrac{\Delta_1^2}{2} + \dfrac{\Delta_2^2}{2}\right)$	$A \doteq \omega\sqrt{LC}\left(\dfrac{\Delta_1}{2} + \dfrac{\Delta_2}{2}\right) = \dfrac{R}{2}\sqrt{\dfrac{C}{L}} + \dfrac{G}{2}\sqrt{\dfrac{L}{C}}$
$B = \sqrt{\dfrac{R\omega C}{2}[\sqrt{(1+\Delta_1^2)(1+\Delta_2^2)} + \Delta_1 - \Delta_2]}$	$B = \omega\sqrt{LC}\sqrt{\tfrac{1}{2}[\sqrt{(1+\Delta_1^2)(1+\Delta_2^2)} - \Delta_1\Delta_2 + 1]}$
$B \doteq \sqrt{\dfrac{R\omega C}{2}}\left(1 + \Delta_1 - \Delta_2 + \dfrac{\Delta_1^2}{2} + \dfrac{\Delta_2^2}{2}\right)$	$B \doteq \omega\sqrt{LC}\,[1 + \tfrac{1}{8}(\Delta_1 - \Delta_2)^2]$
$r = \sqrt{\dfrac{R}{2\omega C}\left(\sqrt{\dfrac{1+\Delta_1^2}{1+\Delta_2^2}} + \dfrac{\Delta_1 + \Delta_2}{1+\Delta_2^2}\right)}$	$r = \sqrt{\dfrac{L}{C}}\sqrt{\dfrac{1}{2}\left(\sqrt{\dfrac{1+\Delta_1^2}{1+\Delta_2^2}} + \dfrac{1+\Delta_1\Delta_2}{1+\Delta_2^2}\right)}$
$r \doteq \sqrt{\dfrac{R}{2\omega C}}\left(1 + \Delta_1 + \Delta_2 + \dfrac{\Delta_1^2}{2} - \dfrac{\Delta_2^2}{2}\right)$	$r \doteq \sqrt{\dfrac{L}{C}}\sqrt{1 + \dfrac{\Delta_1^2}{4} - \dfrac{3\Delta_2^2}{4} + \dfrac{\Delta_1\Delta_2}{2}}$
$x = \sqrt{\dfrac{R}{2\omega C}\left(\sqrt{\dfrac{1+\Delta_1^2}{1+\Delta_2^2}} - \dfrac{\Delta_1 + \Delta_2}{1+\Delta_2^2}\right)}$	$x = \sqrt{\dfrac{L}{C}}\sqrt{\dfrac{1}{2}\left(\sqrt{\dfrac{1+\Delta_1^2}{1+\Delta_2^2}} - \dfrac{1+\Delta_1\Delta_2}{1+\Delta_2^2}\right)}$
$x \doteq \sqrt{\dfrac{R}{2\omega C}}\left(1 - \Delta_1 - \Delta_2 + \dfrac{\Delta_1^2}{2} - \dfrac{\Delta_2^2}{2}\right)$	$x \doteq \sqrt{\dfrac{L}{C}}\left(\dfrac{\Delta_1 - \Delta_2}{2}\right)$

Courtesy of Leon T. Wilson, Coles Signal Laboratory.

TRANSMISSION PRACTICE

10. Ideal and Distortionless Transmission. To meet most[1] engineering requirements transmission is acceptable if distortionless, that is, if all frequency components are transmitted without amplitude or phase-shift discrimination. This would be the case if the attenuation were constant at all frequencies and if B increased linearly with frequency. Practical transmission lines, especially in cable, discriminate against the higher-frequency components by attenuating them more and causing them to arrive later than the lower-frequency components. Compensation can be supplied for these discriminations in the receiving amplifier-stage networks by filters which will favor amplification of the necessary higher frequencies and by delay circuits which will hold back the low[2] frequencies until the arrival of the highs. This technique, especially as applied to the critical transmission problems of long submarine cables, is called *shaping*.[3]

Subjective sensations connected with voice transmission make that form of wire communication more difficult to assess against an ideal. Yet telegraphy's demands often prove to be more exacting than its low signaling speeds (say 22.5 to 100 cps) would imply when compared with a voice mid-band frequency of 1,000 cps; for voice-frequency content in the vicinity of, say, 3,000 cps might be attenuated yet the circuit remain commercially acceptable, whereas the same frequency discrimination in a voice band assigned to telegraphy might render a portion of the spectrum below 3,000 cps useless. A telephone plant consisting of a variety of loaded, nonloaded, and carrier circuits of differing characteristics will transmit speech (200 to 3,500 cps) without impairment; yet the transmission of high-speed [750 bits per second (bps)] digital information (such as signals, data, switching, ringing, telegraphs) will be affected[4] unless confined to the portion of a (voice-frequency) band lying between 700 and 1,700 cps.

Ideal or *perfect transmission* has two additional attributes beyond that of being distortionless: no loss of energy (zero attenuation constant A) and no time displacement or propagation time (zero phase constant B). The first requirement has lost its importance, except for critical relationships of received power to received noise, ever since amplifiers have made it possible to compensate for all power losses on lines and in apparatus. Low propagation time assumes importance in certain duplexed applications: in telephony, in permitting speakers at circuit ends to conduct conversation naturally (Chap. 28, Sec. 32); and in modern forms of telegraphy, in providing for the automatic correction of transmission errors revealed by *error-checking* devices (absolute delay; see Chap. 24, Sec. 71).

Ideal transmission also is not interfered with by external disturbance, nor does it introduce disturbance into other circuits.

Telegraph engineers customarily apply other criteria than transmission accuracy in judging a circuit to meet the "ideal": (1) its total attenuation (that is, its received-signal level) should be such that with standard amplifying repeater equipment it can be joined, end-on, to other sections of line to make up circuits of any required length; (2) it must not have made extraordinary demands as to repeater spacing (in miles); (3) its transmission band should be of such width as to accommodate a standard number of channels without quality deterioration at frequency extremes; (4) its electrical influence must not be felt in adjoining circuits or by radio transmission from its wires or telegraph instruments (Sec. 17).

[1] Sunde, E. D., Theoretical Fundamentals of Pulse Transmission (bibliog.), *Bell System Tech. J.*, **33**, 721, 987, 1954.

[2] Mead, S. P., Phase Distortion and Phase Distortion Correction, *Bell System Tech. J.*, **7**, 195, 1928. Cannon, W. D., Delay Distortion Correction, *WUTR*, **10**, 58, 1956.

[3] Curtis, A. M., Signal Shaping for Submarine Cables, *Bell Lab. Record*, **6**, 237, 1928. Cramer, C. H., Some Modern Techniques in Ocean Cable Telegraphy, *Trans. AIEE*, **66**, 765, 1947. Palmer, W. H., "Duplex Balancing—Non-loaded Submarine Cables," Western Union Eng. Memo. 545, 1949. Wilder, H. F., Electronic Regeneration of Teleprinter Signals, *WUTR*, **5**, 121, 1951.

[4] Horton, A. W., Jr., and H. E. Vaughan, Transmission of Digital Information over Telephone Circuits, *Bell System Tech. J.*, **34**, 511, 1955.

11. Arrival Curves. Expositions are contained in Chap. 1 on the subjects of harmonic functions, the frequency spectra of square-top (rectangular) and sine-wave forms, and the application of Fourier analysis and coefficients to complex wave forms, all as applied to distortionless transmission. In the use of modern carrier techniques the usual received wave shape is that of complex sine-wave modulations (either a-m or f-m) of an a-c carrier; however, the nature of an arrival curve (the plot of received voltage, current, or power vs. time) and the mechanism for shaping it are most easily understood if the given applied sending emf is rectangular.

The perpendicular (abrupt) transition from *space* (S) to *mark* (M) or from mark to space conditions (the effect is the same) contains all frequencies from zero to infinity; and it determines the arrival curve which, but for the frequency nonlinearity of the attenuation A and phase-shift B components of the propagation constant, would also be perpendicular. But if, as customary, the higher frequencies are attenuated or delayed in transmission, the tendency is to round off the corners of the wave in a definite manner, so that a more or less rounded signal is the result. Such rounding off is not objectionable if not carried too far. Since the higher frequencies are the most difficult to transmit, the ideal curve is one which, while giving acceptable signals for the type of telegraph system being used, requires the transmission of the smallest possible amount of the higher frequencies.

The Fourier series for the rectangular perpendicular curve shows the diminishing importance of harmonics of a transmitted frequency beyond the fifth in building up a simulation of the rectangular shape or, what is the same thing, in specifying the amplitude-frequency spectrum of the received wave form. However, in wire-telegraph transmission, even that involving f-m carrier, the required band width is only 2.3 times the fundamental signaling frequency;[1] and in the critical operations of ocean cables it may be as little as $1.65f$ for operating relays.[2]

The arrival curve of an abrupt space-mark sending transition is shown in Fig. 3. The system is determined experimentally or by design[3] to be capable of transmitting frequencies from zero to $\omega_0 = 2\pi f_0$.

$$E_r = E_s e^{-A_0} \cdot f(t)$$

$$f(t) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\omega_0} \frac{\sin \omega t}{\omega} \, d\omega, \text{ which is the function plotted in Fig. 3}$$

Time t, which, as noted above, occurs $B/2\pi f$ sec after application of E_s, marks the beginning of the wave-front slope. The *interval of effective rise* of current $T = \pi/\omega_0 = 1/(2f_0)$. The *effective traffic capacity* of the line in cps $= f_0 = 1/(2T)$; in bps $= 2f_0 = 1/T$.

Where $T = 0.01$ sec, reversals of sending emf can be transmitted up to 100 bits per sec, or 50 cps. The line acts substantially as a band-pass filter. The time required to establish a new carrier amplitude, or a shift in frequency, at the output of such a line or filter is inversely proportional to half the band width[4]—that is, $T = 1/(2f)$.

The arrival curve of a square-top modulated a-c pulse for telegraphy takes twice as long an interval to build up, since in this case $T = 1/f_0$. For equal band widths, d-c telegraphy will theoretically handle twice the information per unit time as double-side-band carrier, and the same amount as SSB transmission.

12. Successions of Pulses on Channels. An understanding of the mechanism of signal transmission over wires is facilitated by considering application at a point A of ungrounded 60-cycle sine-wave a.c. to a metallic pair occupying fixed adjacent pin positions on a pole line joining A with B, 100 miles away. At B, a polar *undulator*

[1] Boughtwood, J. E., Properties of Amplitude and Frequency Modulated Waves, *WUTR*, **3**, 152, 1949.

[2] Milnor, J. W., Submarine Cable Telegraphy, *Trans. AIEE*, **41**, 20, 1922.

[3] Lüschen, F., Modern Communication Systems, *J. IEE*, **71**, 776, 1932.

[4] Bramhall, F. B., Transmission of Business Machine Data over Standard Telegraph Channels, *Trans. AIEE*, **75**, Part 1, 1956 (Paper No. 56-654).

(pen-record galvanometer) traces square-topped alternations produced from the plus-minus contacts of a relay responsive at *B* to *A*'s alternator. The circuit thus demonstrated is capable of being worked at a speed of 60 cps or 120 bits per sec. The bits are the alternations above and below the center line on the undulator tape. They may be given separate identities of mark (M), space (S), to represent the two sine-wave polarities *plus* and *minus* going out from *A*. *B*'s record of the run is therefore . . . MSMSMSMSM . . . .

If *A* were equipped with an electronic *pole changer* that would reverse the alternator polarity every cycle, *B*'s record would read . . . MSSMMSSMMSSMMSSMM . . . . So far as the line and undulator are concerned they would be working momentarily at

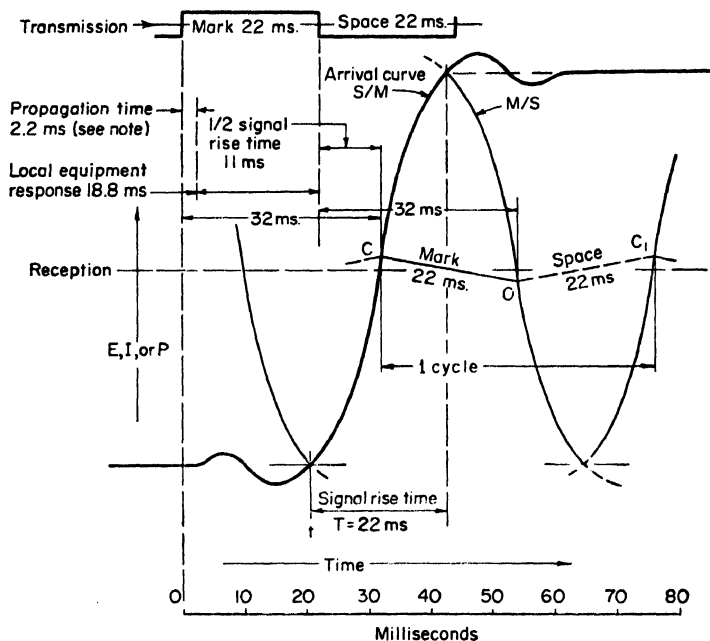



FIG. 3. Arrival curves of 22-millisec S-M and M-S reversals. Note: propagation time, per Table 1, column 5, for 150-mile carrier section on aerial pair is 0.8 millisec; load 50% for carrier repeaters; add 1 millisec for vehicle band equipment—total 2.2 millisec. Local equipment response time additional.

30 cps, or half the working frequency. But the number of intelligence bits per second transmitted remains at 120. So far as the capability of the line to handle signals is concerned it is still a 60-cps line because it must be prepared at any time for *A* to switch out his pole changer and revert to the first pattern.

By synchronously "keying the field" of *A*'s alternator, any combination of *M/S* signals could be sent, including the text of telegrams. A different means of modulation is ordinarily used, in accordance with a teleprinter code, but whatever the means, *B*'s record in handling information would consist of seemingly random combinations of single amid multiple pulses . . . SMMSMSSSSSMMMSMMSMSSSSMSS . . . . . The line has no function other than to transmit faithfully on a time scale each and every *transition* S/M and M/S. Between MM or between SS, the line has no function to perform. The fastest sequences, MSM and SMS, are the ones which tax the capacity of the line to perform; hence a.c. in cycles per second at half the bits-per-second rate tests the line for capacity performance.

13. Characteristic Distortion. The behavior of a polar-operated line and its equipment to changes of polarity is such that the poled arrival curves, upward and downward, normally assume the same shape and a symmetrical position about a horizontal mid-axis, as shown in Fig. 3. The receiver, or its line relay, is presumed to overcome the reverse strength of its biasing and close at a point C on the upward curve just above the axis and to submit to the strength of its biasing and open at a point O on the downward curve about an equal distance below the axis. The horizontal distance between C and O measures the length of the marking (M) pulse initiated by upward-arrival curve S/M; the distance from O to C_1, the corresponding point on the successive S/M curve, is the length of the spacing (S) pulse initiated by downward-arrival curve M/S. Under conditions of symmetry, as would obtain if square-topped a.c. were being transmitted, $C - O = O - C_1$.

Operation would become asymmetrical if, in effect, the axis, carrying the receiver operating points, were raised or lowered, causing them to fall asymmetrically on the arrival curves. This would be the case, for example, if an S/M transition should start from the steady-state level and an M/S transition should take effect before the entire upward interval T had elapsed (that is, during a changing-current transition). The axis, in effect, would be raised and $C - O$ would be shortened to the benefit of $O - C_1$, producing *positive characteristic distortion. Negative characteristic distortion* could similarly be created by virtual lowering of the axis by delaying reversal until the overshoot transient point had been reached. Anything which causes the axis to float up and down the arrival curves will cause characteristic distortion, *i.e.*, the type of distortion stemming from line characteristics (see Chap. 24, Sec. 41).

The principal cause is the occurrence of a succession of like-polarity bits, tending to alter the usual points on arrival curves where reversals take effect.

It will be noted from Fig. 3 and the discussion of Sec. 11 that the complete arrival curve includes tailing, or holdover, transients which are ignored when the impulses of opposite polarity are spaced as closely together as the rising portion of the curve permits. In case an S/M transition is followed immediately, not by an M/S reversal but by another M pulse, the holdover transient of the first pulse will assume additive importance, and a succession like MMMM (or SSSS) will compound the effect, "charging" or "blocking" the line to a polarized bias or permitting the establishment of d-c flow. Under ordinary land-line operation this effect does not require mitigation but demands that account be taken of it in measuring components of the distortion characteristic of the response, "memory," and nonlinearity of the line parameters and terminal apparatus. Bramhall[1] calls it "previous-history" distortion, measures its departures from the normal crossovers of Fig. 3, and notes that good engineering practice holds characteristic distortion within ± 1.0 millisec per carrier section of line.

Since the line characteristics produce normally symmetrical polar arrival curves, any characteristic distortion effects will be the same for SS as for MM, for SSS as for MMM, etc. Since the effects are a function of line parameters, they do not tend to vary from day to day except as temperature and moisture change the line parameters.

In submarine-cable practice, holdover transients possess sufficient comparative values to necessitate measures[2] to prevent the frequently occurring SSSSS signals (the *idling* channel condition of a multiplex) or the less frequently occurring MMMMM (*rub-out* signals) from biasing the cable. This is accomplished by making corresponding polarity changes in the transmitters and printers at the circuit ends so that such successions will be split between the second and third pulses and be sent to line in the form $+ + - - -$ or $- - + + +$.

14. Bias Distortion. Referring to Fig. 3, bias distortion, like characteristic distortion, has effects upon the relative lengths of $C - O$ (M) and $O - C_1$ (S) which may be visualized by a raising or lowering of the axis. In the case of bias distortion, however, the causes are due mainly to unequal M and S voltages or line current, ground potentials, or unintentionally biased relays—causes which are watched and corrected by

[1] Bramhall, *op. cit.*
[2] Haglund, H. H., and A. W. Breyfogel, Thyratrons in Loaded Cable Multiplex, *WUTR*, **3**, 34, 1949.

alert maintenance and which tend to vary from hour to hour. Bias distortion lengthens all M bits alike and correspondingly shortens all S bits alike, and vice versa.

In *neutral* operation but not in *polar* operation, the line shunt parameters, especially leakage conductance G but also ground capacitance coupling C, produce differently shaped arrival curves for the S/M transition than for the M/S transition, thus varying the relationship between $C - O$ and $O - C_1$ without altering the axis position. Such changes may be brought about by rain or fog.

In other neutral operations the arrival curves remain of identical shape. Changing the relay biasing current (or retractile spring) in effect raises and lowers the axis with effects on relative length of mark and space already described. Varying line current by raising transmission voltages will raise the height of the figure, T remaining constant, thus virtually lowering the axis and requiring a biasing readjustment.

The arrival curves are of identical shape in all polar operations. If the marking current is weak because of impressed voltage or earth current or ground-potential differences, the axis will be raised with respect to the curves and a bias will result unless the axis is lowered to mid-point. Relay bias will have effects similar to unequal poles of emf. Correction lies in relay bias adjustment or equalization of effective polar battery.

The circumstance that identical arrival curves exist in polar operation accounts for its immunity to trouble while circuit constants vary—especially leakage conductance as affected by wet and dry weather.

Bias and characteristic distortion are cumulative; together they are given the name of *systematic* distortion. A third important element in total distortion, called *fortuitous* (Sec. 18), involves external circuit influences from adjacent communication circuits and neighboring power-supply lines.

15. Ground Function in Transmission. In an ideal situation independent circuits would contain no electrical elements in common, no coupling, perfect symmetry or balance, and hence no spurious influences from outside. The proximity of all circuits to the ground, however, definitely precludes perfect symmetry and hence perfect balance. Examples of unbalance already considered are (1) the aspect of characteristic impedance of a transmission line when one of its two conductors is the ground; (2) the function of the ground in producing unequal shapes of arrival curves for the make-and-break of neutral operation; (3) the *proximity effect* of many wires on a pole line in altering the capacitance parameter of grounded circuits. In addition the earth disturbs circuit symmetry by entering into the formation of inductive loops containing the wires, and also by the role of subsoil conductivity in introducing spurious currents from one circuit into the other.

Because some of these aspects of circuit proximities and unbalances are geophysical, the properties and functions of grounds in transmission assume great significance.

The resistance of the path in the earth of a d-c circuit of which the earth constitutes one of the two conductors is zero ohms because the specific resistance of the earth's material,[1] whatever it may be, is made negligible by the earth's size. Thus, if a loop consisting of a transatlantic cable and the earth is measured, all the d-c resistance is accounted for in the copper cable conductor, and zero ohms in the earth as a conductor. The earth in fact even constitutes a d-c shunt of the large steel armor wires in contact with the sea.

Since earth's $R = 0$ the potential drop $IR = 0$. It follows that all perfect grounds are at zero potential. The d-c power loss in the ground $I^2R = 0$.

While good connections to ground[2] are obtainable in water, a perfect ground connection is usually to be secured only with difficulty in soil[3] (*cf.* radio ground systems). To minimize ground resistance, resort is often taken to laying multiple grounds in

[1] Card, R. H., Earth Resistivity and Geological Structure, *Trans. AIEE*, **54**, 1153, 1935.

[2] "Grounds," Bureau of Standards Technical Paper 108. J. G. Biddle Co., "Manual for Ground Resistance Testing," 25-J, 1952. Knowlton, A. E., "Standard Handbook for Electrical Engineers," Sec. 14, pars. 70–73, McGraw-Hill, 1941.

[3] Hausmann, E., "Telegraph Engineering," Van Nostrand, 1922.

parallel. From each ground or each set of grounds, *multiple paths* for current through the subsurface may be postulated, flowing from the ground rods or plates into the zero-resistance earth itself. Potential gradients exist, of course, to support the flow of current along these paths, and may become involved with the potential gradients in the grounds of "foreign" systems (Secs. 17, 18).

Because of the relative magnitude of its reactive component compared with its resistance, the earth as a conductor presents a definite impedance to a.c., of practical importance[1] even at 50 cps; and there will be a corresponding power loss in the ground. The fact that a-c potential gradients can be maintained even in sea water is demonstrable by the use of amplified currents from the electrodes of leader gear, used as probes, to locate submarine cables through their external fields.[2] With increases of frequency, cable-armor wires at the sending end carry an increased proportion of the a.c. (regardless of direction), and the sea water in parallel with them less, causing an increase of resistance (steel having replaced the earth) and a decrease of circuit inductance[3] in the system's second conductor (not too accurately referred to as sea "return").

Protector grounds for communication circuits[4] need be no more elaborate than the familiar water-pipe ground. In a different class are those used where current must be dissipated, such as for central telegraph office grounds for circuits, and grounds for cathodic protection of cables against electrolysis.[5] In the latter type the current must flow in the earth without corroding or polarizing the grounds, over long periods of time and without increasing R. Ground resistance decreases with number of electrodes and their diameter and with the depth of driven ground, and it varies with soil conductivity, which may show seasonal variation. A typical office ground measures 0.1 ohm, of which 90 per cent is concentrated in a radius of 20 ft.[6]

Formerly used methods of measurement of grounds by d-c Wheatstone bridge, involving three grounds taken in series pairs, were open to difficulties due to polarization and foreign currents encountered and have been superseded by use of a-c meggers.[7] In wire telegraphy, measured ground impedance is conveniently to be integrated with that of the ground lead or bus.

Although perfect grounds are not always obtainable, they may be assumed in certain situations. Thus a perfectly grounded rod or plate, from an engineering viewpoint, may be assumed to be at ground, or zero, potential relative to all electrically charged bodies, no matter where located on the surface of the earth. Apart from other influences, a wire connecting two such grounds will be "dead" because of the zero potentials existing at its ends.

Ground Currents. Assuming two faultless grounds at points remotely situated, a measurable current, nevertheless, will often flow in a wire grounded at both points. Such currents may flow from three influences:

1. A more or less steady potential difference[8] which exists day after day and which is best taken care of either (*a*) by a ground corrector that applies a counter emf to circuits containing an abnormal ground potential, capable of following and offsetting it, or (*b*) by running a ground wire to a point free of the abnormality, as a cable sea-earth is sometimes reached 100 miles seaward through bicore cable. This condition is usually associated only with man-made electrical effects like those produced by use of the ground as one conductor of electric-railway power. Here, through peculiarities of the subsoil, electrical-potential strata are set up partially insulated from the bulk of the earth's structure.

[1] Wise, W. H., Propagation of High-frequency Currents in Ground Return Circuits, *Proc. IRE*, **22**, 522, 1934.

[2] Lawton, C. S., Electronic Aids to Ocean Cable Maintenance, *WUTR*, **4**, 22, 1950.

[3] Carson, J. R., and J. J. Gilbert, Transmission Characteristics of the Submarine Cable, *J. Franklin Inst.*, **192**, 705, 1921.

[4] Courtesy of A. Z. Mample, Western Union Telegraph Co.

[5] Blomberg, J. B., and J. F. H. Douglas, Cable Sheath Corrosion—Causes and Mitigation, *Trans. AIEE*, **54**, 382, 1935.

[6] Hausmann, *op. cit.*, p. 280.

[7] J. G. Biddle Co., "Megger Ground Testing Instruments," 1954. Associated Research, "Ground Resistance Testing with Vibroground," 1953.

[8] Frost, P. B., The Leakage of Direct Current and Potential Gradients in the Ground, *P. O. Elec. Eng. J.*, **43**, Part 3, 125, 1950.

2. A fluctuating-polarity emf generated dynamo-fashion in the line by moving terrestrial electromagnetic fields—in extreme cases definitely associated with *magnetic storms* and their manifestations of aurora. These are often called *earth currents* or "ECs" but it is not necessary for engineers to assume that their manifestation is due to an actual upsetting of the earth's normal equipotential surface. Mitigation consists in replacing the earth with a parallel wire as the mate to the other conductor; the induced emfs will then be the same and in phase in the two conductors and will cancel. Because of their slow roll (say, 10 to 20 cycles per hr) even a grounded circuit like a submarine cable will work through the disturbance if its (practically d-c) characteristics are separated from the signal currents by employing transformer coupling between cable and amplifier.[1]

3. *Static*, of consequence to wire operations whenever its frequency components reach down into the telegraph signaling spectrum, and whenever its cumulative power at a communication terminal is high enough to break across the tolerable noise-signal ratio. The earth's function in such a power transfer is the same as in radio, the wire (more often submarine cable) line acting as a receiving wave antenna.[2] Lightning has effects on transmission of a higher order of magnitude than static.

Cross Talk Caused by Ground. One way in which the earth disturbs symmetry between communication circuits is as follows: If wires *ab* comprise working circuit 1, and common conductor *a* is used to make up *ac*, *ad*, *ae*, and working circuits 2, 3, 4, the inductive coupling will be closer, and consequent interference between the several circuits greater, than between, say, *ab*, *cd*, *ef*, *gh*, in separate pairs. The close inductive coupling of any two circuits may be looked upon as though one loop were a continuation of the other in the manner of the primary and secondary of an autotransformer. The use of a common conductor *a* for several circuits describes, of course, a pole line loaded with neutrally operated earthed circuits using a common battery. Such systems were once in universal use in telegraphy, but because coefficients of circuit coupling go up with frequency, intercircuit cross talk became intolerable for telephone[3] use wherever parallelisms existed for appreciable distances, and the use of metallic loops and pairs was employed to remove the inductive-coupling influence of the ground. Modern telegraph circuits, operating in the v-f range, likewise universally use metallic pairs.

Another way the earth enters into the creation of a cross influence of circuits due to their proximities or parallelisms is linked with the fact that multiple current paths exist in the subsoil along which circuit energy reaches the earth as a conductor. It follows that potential gradients exist in the subsoil to support the flow of current along these paths, and that they may become involved with the potential gradients set up by virtue of the presence, in the same body of subsoil, of the ground systems of neighboring communication or power-supply circuits.[4] Coupling currents will be set up in conductive soil between all points of unequal potential, resulting in the introduction of spurious energy into each circuit by the other. This type of ground coupling is the most serious cause of power induction into communication circuits.

16. Power Induction. Extensive parallelisms between power-supply lines and wire-communication circuits are avoided wherever possible to reduce power interference in the communication lines.[5] Cooperative measures by the industries involved to avoid interference has become known as *inductive coordination*,[6] in the pursuit of which the engineers recognize the public's right to both services at minimum over-all cost.

For convenience, inquiries are directed between (1) the lower-frequency effects[7] of

[1] Cramer, C. H., Sunspots and Telegraphy, *WUTR*, **1**, 43, 1947. Haglund, H. H., and A. W. Breyfogel, Thyratrons in Loaded Cable Multiplex, *WUTR*, **3**, 34, January, 1949.

[2] Cramer, C. H., Submerged Repeaters for Long Submarine Telegraph Cables, *WUTR*, **5**, 81, 1951.

[3] Carson, J. R., Wave Propagation in Overhead Wires with Ground Return, *Bell System Tech. J.*, **5**, 539, 1926.

[4] Wise, W. H., Potential Coefficients for Ground Return Circuits, *Bell System Tech. J.*, **27**, 365, 1948. Sunde, E. D., "Earth Conduction Effects in Transmitting Systems," Van Nostrand.

[5] Courtesy of E. A. Cooke, Western Union Telegraph Co.

[6] Pack, R. F., W. H. Harrison, and A. E. Silver, Symposium on Coordination of Power and Telephone Plant, *Trans. AIEE*, **50**, 437, 1931.

[7] Conwell, R. N., and H. S. Warren, Low-frequency Induction, *Trans. AIEE*, **50**, 458, 1931.

60 or 25 cps and their harmonics and (2) the effects of power-line abnormalities of higher frequency[1] due to switching, insulation breakdown, and lightning transients.

As noted in Sec. 15 the effects of all influences are multiplied whenever they involve common ground coupling or potential gradients in the earth.

If a three-phase power circuit is balanced as to its parameters and its loads, the vector sums of the phase voltages and currents to ground are zero. A fault to ground on one phase[2] will produce vectors representing residual voltage and current, no longer zero. Residual currents are equivalent to single-phase currents applied to a circuit composed of the ground and all three power conductors in parallel. The effects on communication lines are due to the physical size and electrical values of the coupled loop, and the magnitude of the currents and potentials in the subsoil. The extraneous voltages which appear are a function of the exposure conditions of the current and voltage of the power line, the fundamental and harmonic frequencies, and the degree of balance in power and communication lines.

During normal power-system operation, telegraph carrier circuits are most affected by harmonics of the power-line fundamental frequency, while d-c telegraph is more susceptible to the fundamental and third harmonic.

Transient voltages at their highest are induced into communication circuits when faults or abnormalities occur on paralleling power lines, frequently because of lightning. Such transients may be sufficient to operate protective equipment on the communication circuits.

Specific remedial measures include reduction of coupling by providing greater separation, improved balance of power-line voltages and current or of metallic communications pairs, transpositions, shield wires, use of cable, resonant shunts, limitation of fault currents, self-restoring arresters on communication circuits, and *neutralizing transformers*.[3] The latter have one primary winding and from 2 to 100 pairs of secondary windings to accommodate the communication circuits.[4] The primary winding is energized by a voltage of such phase and magnitude as to neutralize the voltage induced in the telegraph wires connected to the secondaries, and this may be acquired by direct connection to a low-impedance pilot wire subjected to the same exposure or by indirect connection through tube amplifiers.[5] The primary winding may be tuned to the interfering frequencies to increase its effectiveness and to reduce the cross talk between the communication circuits.

For the treatment of lesser cases such as individual instruments and short patron's loops, tuned resonant shunts are helpful where signaling speeds are 30 cps or less and the interfering frequency is 60 cps or a harmonic thereof.

General measures of control are directed toward minimizing the influence of the power circuits, decreasing the susceptibility of the communication circuits, reducing the coupling between the circuits, and various combinations of these factors.

Economics, reflecting high costs of modern power and communication lines, often dictates acquisition of private rights of way for them, thus reducing parallelisms by the avoidance of highways. Some railroads provide common rights of way, creating long exposures of power and communication lines.[6]

17. Telegraph Cross Talk. The coupled interference of energy carried by one communication wire or circuit with that in a parallel communication wire or circuit is descriptively similar to the power-interference problems discussed in Sec. 16 except that parallelisms and proximities cannot be avoided. In telephony and carrier teleg-

[1] Wills, H. L., and O. B. Blackwell, Noise Frequency Induction, *Trans. AIEE*, **50,** 448, 1931.

[2] Conwell and Warren, *op. cit.*

[3] "Neutralizing Transformers," Eng. Report No. 44; Joint Subcommittee on Development and Research, Edison Electric Institute and Bell Telephone System (E.E.I. Pub. H12).

[4] Courtesy of J. R. Hyneman, Western Union Telegraph Co.

[5] Milnor, J. W., Control of Inductive Interference in Telegraph Systems, *Elec. Eng.*, **59,** Part 2, 469, 1940.

[6] A thorough technical treatment of inductive coordination is to be found in Recommendations 7-7 of the Association of American Railroads, "Discussion of Fundamental Factors Involved in Inductive Coordination and of Remedial Measures Applicable under Various Conditions," and in Recommendations 7-8, "Practical Procedures for Investigating Inductive Effects on Communication Facilities."

raphy such interference is known as cross talk; in d-c telegraphy it is known as cross fire.[1] It is customary to think of both as inclusive of both power and communication interference components. Cross fire, as a separate subject, has gone out with the day of heavy leads of grounded time-division-multiplex circuits.

The functions of unavoidable subsoil current interactions in increasing interference have been sufficiently emphasized; the observations apply equally well to cross talk. In communication parallelisms, partial grounds occur wherever shunt conductance is affected by partial failure of line insulation, inducing cross talk.

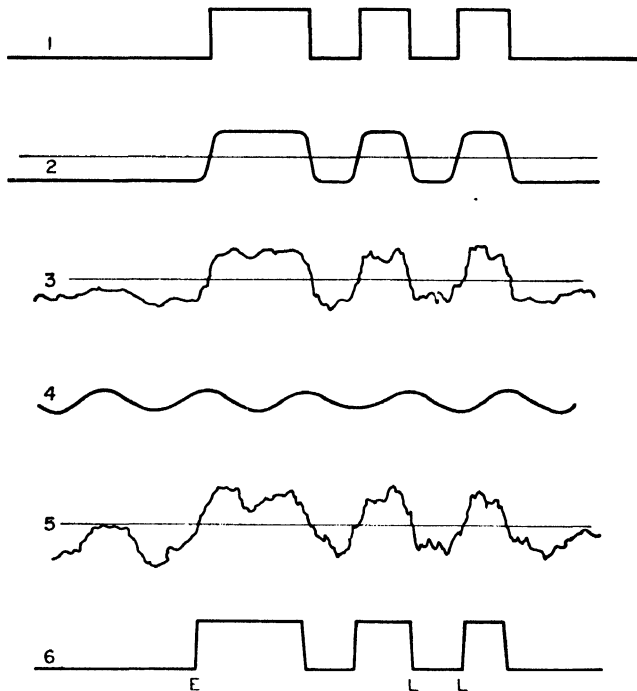



Fig. 4. Effects of interference on transmission. (1) Signal as transmitted. (2) Rounded-off received signals without disturbance, (3) with telegraph cross fire. (4) 60-cycle a-c power interference, superposed (5) on signals and cross-fire (3). (6) Early (E) and late (L) relay displacements due to interference.

Apart from leakage, any two conductors, one of which carries signals, are coupled electrostatically by shunt capacitance directly between them and by series capacitance between each and ground; at the same time they are coupled electromagnetically by the moving field associated with the wave fronts in one conductor inducing an emf in the other. Induction results from both types of coupling. Ordinarily the two effects are measured as a whole and included in cross talk. Figures 4 and 10 show how telegraph signals are distorted by power-supply and telegraphic cross talk.

With spread of the use of carrier frequencies for intercity telegraphy, the telegraph system now operates, as does the telephone system, relatively free of grounded circuits. Aerial and multipair cables and open-wire lines constitute aggregations of metallic loops.

Telegraphic cross talk into telephone circuits is aggravated by any great disparity

[1] Shanck, R. B., Neutralization of Telegraph Crossfire, *Bell System Tech. J.*, **5**, 418, 1926. Gould, K. E., Coupling between Parallel Earth-return Circuits under D-c Transient Conditions, *Trans. AIEE*, **56**, 1159, 1937.

between signaling voltages used on the two systems, so that practice is to limit telegraph signaling current in the telephone plant.

In aerial and underground cable, the pairs are laid throughout each unit length in layers or in bunches, adjacent pairs having different lengths of twist to prevent cross talk. Pairs in the cable center are frequently twisted together to form *quads.* An effort is made in splicing cables to have each pair occupy an approximately average position with relation to the grounded sheath that every other pair occupies. Equality of external induction experienced by each of the wires in a cable pair is enhanced by close spacing; hence pairs in cables are less affected by induction from nearby pairs than are those on pole lines.[1] A lead cable sheath also acts as a substantial (50 to 100 per cent) *electrostatic shield* against the minor electric-field component of power-line interference.[2] The outer conductors themselves also act as a collective shield for the inner wires. Similar external mutual shielding effect may be observed in a pole line as a whole in relation to any pair included.

In open-wire line construction the tendency is away from uniform 12-in. spacing of the 10 pin positions on each crossarm toward 8-, 6-, and even 4-in. separations within each pair and correspondingly increased separations between pairs. Closer spacing is dictated by inductive considerations as carrier frequencies increase; collateral benefit is secured by a decrease in L, which (*cf.* Table 2), although increasing the attenuation constant A (which can be retrieved by terminal or repeater amplifier gain), desirably increases wave-front velocity and decreases the phase shift B.

Equalizing and stabilizing of the two series parameters R and L of the two conductors constituting each pair, and balancing the shunt parameters C and G of all pairs in relation to each other are the desiderata in reducing mutual interference between pairs. It is essential that the two conductors of a pair be of the same material and gage at all points on the line. In stabilizing R, the rule is to solder all terminal joints in the circuit, to make joints in cable by twisting, and in aerial wires to use cold-flow compression-sleeve joints. Jack-terminal machine wrapping techniques are replacing soldering in some applications. Equalizing L involves the equispacing of any loading, continuous or lumped, applied to the line.

Transposition. Shunt-parameter balancing is accomplished through transposing conductors, in the relative pin positions they occupy on the crossarms, as their course continues from pole to pole. For many years such pin-position permutations have been widely used in telephony. With adoption of metallic carriers, *transposition patterns* have assumed equal importance in telegraphy.

A former method of securing approximately equal inductive exposure, used on open-wire grounded telegraph lines, was to route single conductors throughout all pin positions on the line in an attempt to equalize its linkages with all other conductors. Present-day transposition patterns do not follow that principle. With the exception of the two side circuits and phantom of a four-wire phantom group, which follows such an equal-exposure pattern but the use of which for telegraphy has nearly ceased, transpositions are confined to the two conductors of each pair; the pair continues to occupy the same two pin positions throughout the length of line.

Two guiding principles may be recognized,[3] working concurrently: First, it is sought to prevent loss of energy by induction from the pair to all the rest of the wires on the line considered as a group and having a general relationship with the ground. This is accomplished by frequent, more or less constant, interchange, using extra types to be described; the extra pattern for one pair will not be unlike those for all others. Second, it is sought to use a wide selection of different *fundamental types* so that each pair will appear distinctive in relationship to most others, and so that, especially respecting its near neighbors on the line, it will present true relative inductive cancellations or offsets.

The first principle enunciated has direct bearing upon attenuation and noise-power ratios in a pair. If energy can be held within a pair rather than lost to the line as a

[1] Everitt, W. L., "Communication Engineering," 2d ed., Chap. 8, McGraw-Hill, 1937.

[2] Conwell and Warren, *op. cit.* Wills and Blackwell, *op. cit.* Gooding, F. H., and H. B. Slade, Shielding of Communication Cables, *Trans. AIEE,* **74,** Part I, 378, 1955.

[3] Courtesy of T. F. Cofer, Western Union Telegraph Company.

whole it not only will increase received power on the pair but will decrease the totality of noise on that pair received from all pairs on the line as a group.

The second principle focuses attention on linkages between one pair and the others individually, especially neighbors. It works hand in hand with physical layout of crossarms and pin separations and goes even to decisions as to which side of each insulator shall bear each conductor. As frequencies have gone up for carrier operation, the tendency has been to increase crossarm vertical separation from 20 to 24 and 30 in. Conductor separations in a pair have come down from 12 to 6 in., and spacing between adjacent conductors of contiguous pairs on the same crossarm has increased from 12 to 30 in.

Extremes of inductive coupling take place: (1) between pairs mounted adjacently on the same crossarm, for despite the fact that their loops lie in the same horizontal plane, one of the conductors of a pair lies nearer than the other to an interfering pair, producing an unbalanced configuration; (2) between any pair and the one just above or below it on the next crossarm, for coupling occurs by transformer action between the loops, regardless of their balanced configuration. The transposition designer must pay particular attention to such exposures.[1]

Consideration also has to be given to the interference introduced in proximate pairs by waves traveling in opposite directions, leading to a sharp difference between the interference levels of strong near-end (or sending-end) cross talk (where an attenuated received wave is being mixed with strong induction from outgoing signals) and weaker far-end (or receiving-end) cross talk (where both the wanted signals and unwanted noise are traveling and are being attenuated in the same, incoming direction). Near-end cross talk at carrier frequencies is so difficult to cope with in cable that separate directional cables are often employed in four-wire connections to avoid it,[2] while on open wire-line sections, though two-wire carrier is used, a strict frequency-cleavage pattern is employed to separate east and west transmissions. Transpositions may then be designed to conform to frequency separation, and near-end cross talk will be minimized.

FIG. 5. Paired pin positions, "40-wire" line, facing direction of ascending pole numbers. 5-6, 15-16, etc., are pole pairs; others are side pairs. Wire tie positions on insulators indicated. Figures below crossarms denote spacing in inches. Letters with subscript figures above lower crossarms illustrate fundamental and extra transposition patterns employed. (*From Monograph* 2520, *Courtesy Bell System*.)

To illustrate only a few of the expedients employed in transposition engineering, a design specialty, attention will be confined to a single example. In a typical 8-span transposition pattern of 40 wires borne on the 4 crossarms (Fig. 5) of a pole-line trans-

[1] Osborne, H. S., The Design of Transpositions for Parallel Power and Telephone Circuits, *Trans. AIEE*, **37**, 897, 1919. Chapman, A. G., Open-wire Crosstalk, *Bell System Tech. J.*, **13**, 19, 1934.

[2] Searle, G. W., and R. A. Schaefer, Combined Operation of 24-channel Cable Carrier and 12-channel Open-wire Carrier Systems, *Trans. AIEE*, **74**, Part 1, 770, 1955. Moskowitz, S., L. Diven, and L. Feit, Crosstalk Considerations in Time-division Multiplex Systems, *Elec. Commun.*, **28**, 209, 1951.

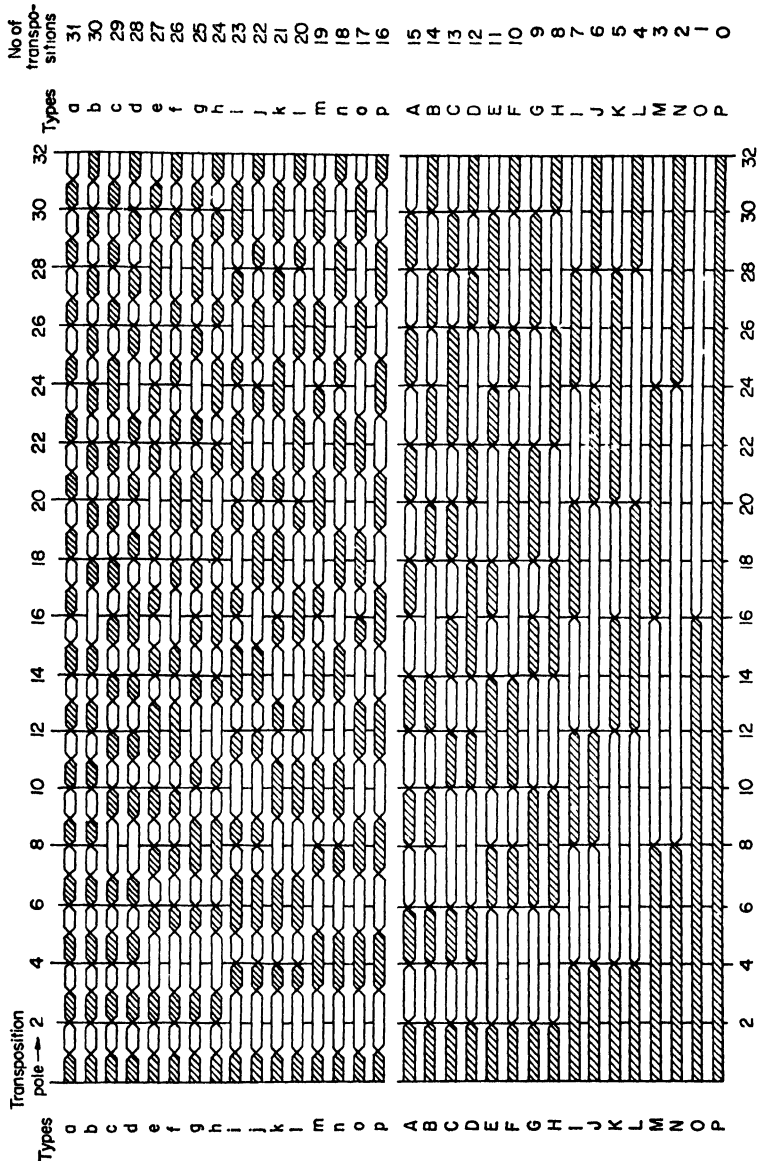



Fig. 6. Fundamental transposition patterns, typical of 8-pole (8-span) units between transposition poles; 32 transposition poles per 256-span section (6.4 miles). (*Courtesy Bell System.*)

position section 6.4 miles in length (257 poles, 256 spans, span length 130 ft, 40 poles per mile), all 20 pairs on the line are subject to possible fundamental transposition of pin positions at least every eighth pole (every 8 spans, or 32 possible transpositions throughout the section length), following one of the 32 types, *a* to *p* and *A* to *P*, of Fig. 6. Thus, pair 13-14, transposed as type *a*, would be transposed at every eighth span, while 27-28, following type *A*, would be transposed at every sixteenth span.

It should be noted that a pair of type *b* would have no relative transposition to a pair conforming to type *a* until the 128th pole (at mid-section) was skipped by the type *b* pair, after which the pairs would be in opposition. This relative opposition itself can be diagramed, as type *O* of Fig. 6. Each combination of any two transposition patterns will be found to have one relative in Fig. 6. The object of *relative transposition* with other pairs is to subject each pair, farther along the line, to counteracting influences in the right proportions after each exposure to induction by parallelism with nearby pairs.

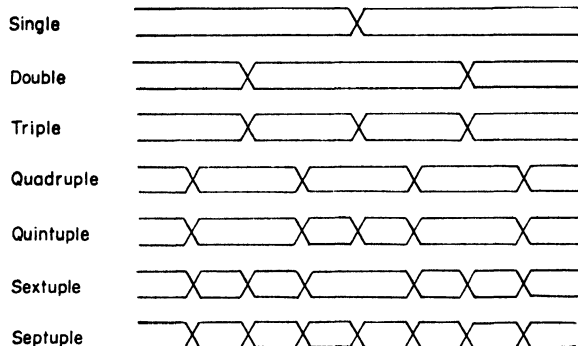



FIG. 7. Extra transposition patterns. Widely used extra transpositions occur within 8-span fundamental transposition intervals. (*Courtesy Bell System.*)

The frequently transposed patterns, like *a* to *f* of Fig. 6, also serve the other purpose of transpositions—to prevent loss of energy from the pair to the line as a whole. As frequencies go up and the point is approached where there is a tendency to radiate energy, transpositions based on the fundamental eight-span pattern prove inadequate. Further refinement takes the form of the extra types of Fig. 7. The *septuple extra* would involve transpositions on every pole or in every span; in practice, the employment of extras rarely needs to go beyond the *quintuple extra* type.

The transposition configuration of one section of a line is not necessarily that of adjacent or successive sections. *Transitional patterns* are employed, and separate consideration has to be given every portion of open-wire line marked off by discontinuities such as points of juncture with branch circuits, impedance mismatches, pole-line configuration changes, points of power-induction exposure, and circuit terminations.

The function of transpositions is to prevent energy losses and to equalize exposures; exposures are functions of moving electromagnetic fields and therefore of applied emfs, attenuation *A*, and phase shift *B*, in the interfering and interfered-with conductors, bringing with them a varying rather than a constant influence as a wave proceeds along a line. These variables must be taken into account.[1]

The idealized illustration chosen makes no allowance for difficulties encountered in actual lines,[2] where unequal pole spacing, variable sag in line wires, and nonuniform spacing of crossarms, pins, and wire ties require attention of the transposition designer. Many older transposition schemes were worked out on 8-mile instead of 6.4-mile sections, and on 10-span instead of 8-span fundamental patterns, calling for compromises in the extra types of transpositions. The cumulative effect of these variables does not present insuperable transposition difficulties for lines and pairs carrying 30-kc signals but is important enough, with 150-kc carriers, to dictate resetting of poles, resort to the use of accurately placed span-transposition brackets, or rebuilding of lines.

To complete consideration of interference, note should be taken of measures adopted to eliminate the interference of wire-communication circuits and apparatus with radio receivers,[3] especially in a-m broadcasting. Telegraph wires and telegraph power-

[1] Rentrop, E., The Type-01 Transposition System, *Bell Lab. Record*, **33**, 45, 1955.

[2] Cofer, T. F., Printing Telegraphy and Inductive Disturbances, *WUTR*, **1**, 48, 1947.

[3] Newell, E. L., T. F. Cofer, and F. H. Cusask, The Suppression of Radio Interference from Telegraph Equipment, *WUTR*, **4**, 137, 1950. Gill, A. J., and S. Whitehead, Electrical Interference with Radio Reception, *J. IEE*, **83**, 345, 1938.

supply conductors act as antennas for sporadic r-f emissions from sparking-relay contacts, teleprinters, and other apparatus in telegraph and customers' offices. Of 25 types of corrective networks employed, three representative types are shown in Fig. 8. The spark killer A is most widely used and is placed as close to the contacts as possible with short leads; a contact-protecting filter B is used to eliminate interference in teleprinter applications; a filter C eliminates interference of generators via power-supply lines.

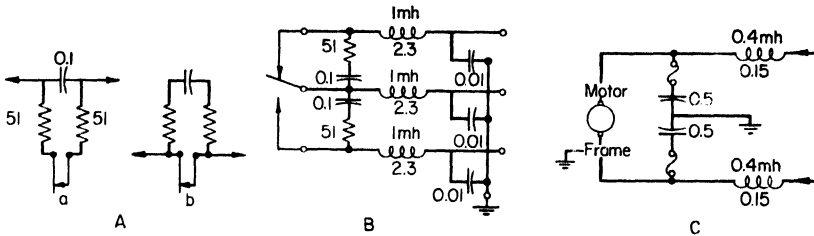



FIG. 8. Measures for prevention of interference to radio: (A) contact-shunting spark killers; (B) contact-protecting telegraph interference filter; (C) power-line interference filter.

18. Fortuitous Distortion. The accumulated effect of all the elements which have been described as contributing to induced interference, or noise, in communication circuits is called fortuitous distortion. By interaction of noise power N with the arrival curves of residual signal power P, the curves become distorted in shape; the points where they cross the horizontal line which defines the operating level on the arrival curves are displaced an amount measurable in milliseconds or microseconds (or expressed as a percentage of the signal length) from datum points established by the receiver's *sampling* or gating circuit. A representatively safe signal displacement in the first carrier section of a line, due to fortuitous distortion,[1] is 5 per cent of 22 millisec, or 1.1 millisec.

The ratio N/P is called the noise ratio; P/N is the signal-to-noise ratio; the percentage of noise is $100N/(P+N)$. Noise in telegraph circuits, and its influence upon fortuitous distortion, is characterized by long intervals during which its background effects remain within a constant range, punctuated by erratic jumps to higher values. The peaks determine the circuit *margin*.

FIG. 9. Differential duplex, one-wire and ground, polar. If Z of the artificial line AL is made equal to Z of the main line beyond P, each station's relay will be unresponsive to its own key but will remain responsive to the distant station's key.

A source of interference appearing as a physical-circuit noise component but not falling into any of the classifications enumerated so far is that introduced into the receiving end of a duplexed circuit by the transmitting going on at the same end (see Fig. 9). This effect results from inexact impedance matching (called *balancing*) of the artificial line AL, with the main line ML at all fundamental and harmonic frequencies included in the transmitted wave form (usually square-topped d.c.). The

[1] Bramhall, F. B., Transmission of Business Machine Data over Standard Telegraph Channels, *WUTR*, **10**, 89, 1956.

mismatch causes points P and Q to be at different potentials instead of at the same potential when the battery of either pole is applied to the apex X; the receiving relay R_a, in consequence, senses spurious locally generated interference with the wanted received signal coming in over the line conductor at point P and taking a double path to the second conductor through X and Q.

Because of the similarity of its effects and difficulty of separating them from inductive disturbances in measurement, duplex unbalance is usually included within the category of fortuitous interference.

Trends in the communication industry are away from duplexed and other electrically balanced telegraph circuits of all kinds, partly because of the expense and lost time associated with keeping them sufficiently in balance to prevent self-induced circuit noise. Among such balanced circuits are derived *composite* grounded d-c telegraph circuits on each of the two conductors constituting a metallic pair in telephone use; *phantom* pairs for telegraph carriers derived by mid-tapping the line-coupling coils of two ungrounded *side pairs;* and *simplex* one-way telegraph circuits one of whose conductors is the ground and the other a metallic telephone pair the line-coupling coils of which are mid-tapped. Four-wire connections for two-way carrier are preferred to frequency-divided two-way circuits in multipair cable. The availability of undersea amplifiers has accelerated the abandonment of duplexed submarine cables except on single-cable routes. Radiotelegraph employs one-way circuits.

Hence there has been a growing avoidance of electrically balanced circuits. Even when full two-way teleprinter duplexes are set up, resort is often had to "half duplexing," or back-and-forth operation alternately. In the balanced configuration of Fig. 9 this is accomplished by introducing a differential control relay[1] at the apex in series with the relay shown, the function of which, in connection with a neutral relay and a differential break relay, is to deny simultaneous two-way use of the line to the operators at the two ends without denying to either of them, by manipulation of his sending keyboard, the ability to break up and hence to stop the other's transmission. Such an arrangement is less critical as to maintenance of duplex balance than one designed to afford continuous two-way use.

19. Measurement of Distortion. Several distortion-measuring instruments and methods have found acceptance. The device most widely and pragmatically used is the teleprinter itself with its incorporated ranging scale.[2] More accurate findings may be made with electronic testing equipment.

In Fig. 10, S-M and M-S transitions are represented by arrival curves which show the effects of distortion. The output of the main line relay is seen to be broken up at O and C. Such a signal, though it has indefinite boundaries, is useful, in the middle portion MN, to determine definitely the polarity of the signal. Multiplex receiving brushes (or, in the case of start-stop teleprinters, the receiving cams) are arranged to sense the polarity of the signal only in the middle of the useful portion PQ. That much (one-half) of the pulse length is adequate for the local operation of relays and printers, but for onward transmission into another line or into a carrier section the full-length pulse PR is *regenerated* by locking up, in or after the interval QR, an extension of the signal sampled during the interval PQ.

Start-stop teleprinters are equipped with a *ranging arm* and associated scale, attached to the starting mechanism of the receiving system, to refine the orientation of the receiving cams to the middle portions of signals.[3] With this arm the starting latch may be rotated with respect to the position of the No. 1 receiving cam, by this means determining whether the mechanism will be actuated by the leading edge, center, or trailing edge of the signal. The total travel of the range arm is made 1.2 times the length of one signal pulse; so the scale is graduated into 120 points, 100 points range being ideal for perfect signal and perfect mechanism. Wear of parts and

[1] Pender, H., and K. McIwain, "Electrical Engineers' Handbook," 4th ed., pp. 18–23, Wiley, 1950.

[2] Rea, W. T., Effect of Telegraph Distortion on the Margins of Operation of Start-Stop Receivers, *Bell System Tech. J.*, **23**, 207, 1944.

[3] Cofer, T. F., Printing Telegraphy and Inductive Disturbances, *WUTR*, **1**, 48, 1947.

instabilities will bring the range down to 80 before the line is connected to the teleprinter, not necessarily centered at 60, mid-point on the scale.

The *ranging test* is regularly performed before the circuit is opened for traffic, and at other times as errors in received copy reveal loss of range. By rotation of the range arm to its two extreme positions, the points on the leading and trailing edge of the signal are determined where false signals appear, and the scale is read in each case. The range arm is then moved to and locked at the point on the scale which is the arithmetical

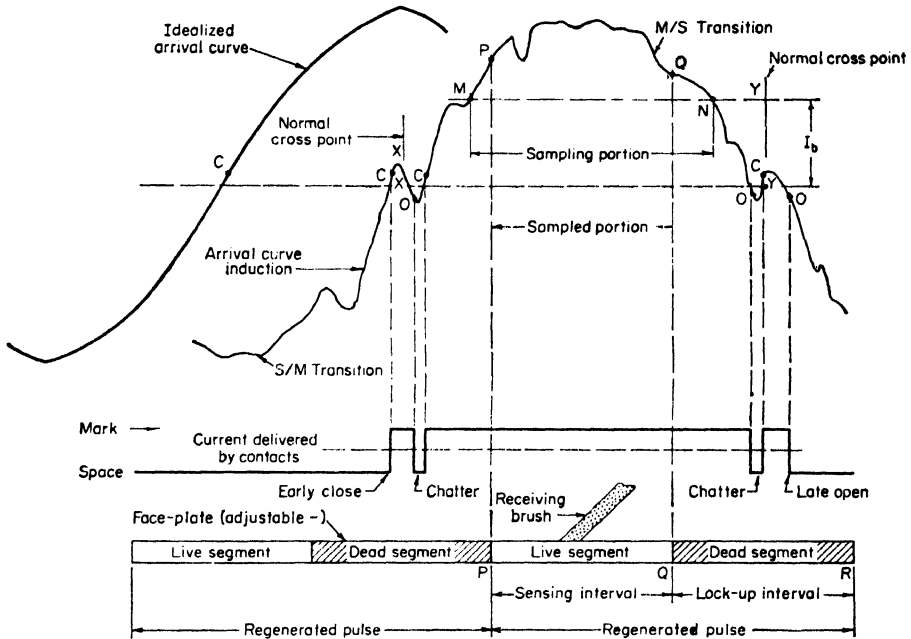



FIG. 10. Arrival curve, with induction, showing response of relay contacts and sampling and regenerative action of receiving brushes and faceplate. The break-over current I_b is that required in the relay windings to prevent contact tongue from breaking contact (chatter) due to inductive disturbances. The faceplate segments can be oriented as desired to the middle of incoming signals.

average of the upper and lower limit readings. The *range* is stated as the difference between those readings. It will remain valid until some change in transmission occurs to narrow it or to shift it off center, in which event a new range must be taken and its center determined.

More accurate measuring methods[1] are concerned with the correction of distortion due to *envelope delay*, the first derivative of the phase constant B with respect to $2\pi f$, over the significant range of frequencies encountered in facsimile or v-f transmission. A number of representative papers have been published on the measurement of distortion.[2]

[1] Cannon, W. D., An Envelope Delay Measuring Instrument in the Audio-frequency Range, *Trans. AIEE*, **74**, Part 1, 710, 1955. Cannon, W. D., Delay Distortion Correction, *Trans. AIEE*, **75**, Part 1, 55, 1956.

[2] Cramer, C. H., Some Modern Techniques in Ocean Cable Telegraphy, *Trans. AIEE*, **66**, 494, 1947. Erikson, G. L., A Telegraph Signal Analyzer, *Trans. AIEE*, **67**, Part 1, 380, 1948. Wheeler, L. K., and R. S. Tissington, An Electronic Distortion Measuring Set for Start-Stop Telegraph Signals, *P. O. Elec. Eng. J.*, **43**, Part 1, 18, 1950. Wilder, H. F., A Teleprinter Signal Bias Meter, *Trans. AIEE*, **70**, Part 2, 703, 1951. Cory, S. I., A New Portable Telegraph Transmission Measuring Set, *Trans. AIEE*, **73**, Part 1, 59, 1954. Wheeler, L. K., and A. C. Frost, A Telegraph Distortion Analyzer, *P. O. Elec. Eng. J.*, **47**, Part 1, 5, 1954. Cannon, W. D., A Portable Telegraph Bias and Distortion Measuring Instrument, *Trans. AIEE*, **74**, Part 1, 401, 1955.

Because of many differences in their causes, characteristics, and simultaneity of occurrence, the contributions of characteristic, bias, and fortuitous distortion are not directly additive as to sign, with respect to either a single line section or a succession of them. Within a single section, generally speaking, bias which is constant is additive to the rms of characteristic and fortuitous components, which are erratic. From section to section, characteristic distortion tends to be cumulative, since unfavorable signal combinations will adversely influence every section alike, while bias and fortuitous elements will assume rms relationships. For 100-wpm teleprinters working over a single carrier section, good engineering practice will specify limits of 1 millisec signal variation (7 per cent of the 13.3-millisec signal length) for characteristic distortion, and 5 per cent each for bias and noise; the total is not 17 per cent but less than 13 per cent because of rms relationships.[1]

Rms distortion losses up to 13 per cent of the length of a signal, disclosed by a teleprinter on its range scale or otherwise measured, are well within the limits of safety from errors. Variations exceeding 35 per cent will cause errors. It is the practice to regenerate signals completely whenever, after three or four tandem carrier interconnections, the variations exceed 30 per cent of full-length signal.

20. Accuracy of Transmission. Perfect transmission requires that all the type bars of all the receiving teleprinters of all the channels on a circuit shall at all times, without error, reproduce the continuing functions of the sending keyboards. While this ideal is never long sustained in practice, it is possible to assign and sustain any required standards of lines up to but not including the teleprinters, provided that no limit is set by economics, that the physical integrity of the circuit constants is preserved, and that the plant is protected from acts of God.

The Bell System employs telegraph *transmission coefficients*[2] to epitomize distortion conditions affecting transmission accuracy in each component of its end-to-end interconnections. Such components comprise sections of line, by types; loops, switchboards, and repeaters. The coefficients run from 0.0 to 10.0; like decibel attenuation figures, they are additive, component by component, so that a circuit total coefficient may be directly ascertained as the sum of the coefficients of its parts. The highest-quality lines,[3] such as a v-f carrier teletypewriter section, will therefore have a low coefficient, like 1.5, while an inferior grounded half-duplex line to an outlying point will be weighted 4.0 or 5.0. An entire circuit is considered satisfactory for commercial service if the coefficient $\leqq 10.0$. *Regenerators* are inserted to divide circuits with a coefficient > 10.0; working limit between regenerators is 7.0.

Line coefficients reflect all causes of distortion and degrading, but not failures, disconnects, accidents, and interruptions. They are numerically the product of the rms values of the several distortion influences multiplied by a conversion factor determined by normal-law distribution of the probability that a limit of 35 per cent signal distortion (which produces errors) will be exceeded in one case out of an arbitrary number of signal transitions.

Practically, coefficients are related both (1) to a maximum checkable distortion under normal operating conditions, expressed as a percentage of signal length (*cf.* Sec. 19) and determined over a representative 5-min run, and (2) to the number of characters which may be presumed to be handled with complete accuracy per single errored letter. These relationships are shown in Table 3, adapted from published graphs.[4] The most significant point is the coefficient of 10.0, beyond which a circuit requires attention to render it generally acceptable. Here the average transmission develops only one wrong character in 44,000—equivalent to one error in 12 pages of single-spaced typing, letterhead size, or one error in 2 hr at 60 wpm. All lower coeffi-

[1] Bramhall, F. B., Transmission of Business Machine Data over Standard Telegraph Channels, *Trans. AIEE*, **75**, Part 1, 1956 (Paper No. 56-654).
[2] Duncan, J. A., R. D. Parker, and R. E. Pierce, Telegraphy in the Bell System, *Trans. AIEE*, **63**, 1032, 1944.
[3] Cory, S. I., Telegraph Transmission Coefficients, *Bell Lab. Record*, **33**, 11, 1955.
[4] *Ibid.*

cients represent better accuracy; the median is 6.5.[1] The standard mean coefficient for private-line telegraph service is 6.5, with an accuracy performance of one error in 8,000,000 characters; the mean for TWX (teletypewriter exchange interconnections, subscriber to subscriber) is 3.5, because of the prevalence of short hauls and simpler circuitry than in private wire networks; 35 per cent of TWX connections have coefficients of less than 2.0, pointing to line accuracy far beyond 100 billion characters per error due to line transmission (see Chap. 24, Sec. 56).

Table 3. Design Correlation between Telegraph Transmission Coefficients, Distortion, and Circuit Accuracy—Intercity Lines

Transmission coefficient	Private-line services, % of circuits having coefficients exceeding those tabulated	No. characters per error due to circuit	Design max, short-run distortion, %
12	11	7,000	
11	17	11,000	
10*	23	44,000	18.0
9	30	110,000	17.2
8	38	500,000	16.3
7†	46	2,500,000	15.3
6.5‡	50	8,000,000	14.6
6	54	25,000,000	14.0
5	63	900,000,000	12.6
4	74	100,000,000,000	11.2
3.5¶	79	....	10.4
3	83	....	9.4
2	91	....	7.4
1	96	....	5.0
0	100		

Source (adaptation from graphs): Cory, S. I., Telegraph Transmission Coefficients, *Bell Lab. Record*, **33**, 11, 1955.
* Upper coefficient design limit, over-all circuit.
† Upper coefficient design limit, regenerated section of line.
‡ Mean coefficient for private-wire services.
¶ Mean coefficient for teletypewriter exchange connections.

PLANT RELIABILITY

21. Circuit Discontinuities. A practical consideration in the operation of lines is the occasional occurrence of conductor discontinuity or high resistance, radically affecting R, and of short-circuiting of line conductors, affecting G in some section of a circuit. These are known as interruptions or faults, and they take many forms. Discontinuities due to wire breaks are comparatively easy to locate and remedy by wire or pair substitution in the affected section. High-resistance faults due to poor or corroded joints in conductors or joining of wires to equipment, giving rise to intermittent faults, are avoided by use of proper construction and maintenance techniques[2] and tools, producing soldered joints, cold-flow "welds," and other measures[3] to prevent joint oxidation. The improper use of soldering flux can itself contribute to line faults.[4]

A modern 12-channel 24-repeater carrier, any one of whose v-f channels may be modulated by telegraph signals, contains approximately 6,000 soldered connections

[1] Gryb, R. M., Error Checking with Particular Reference to Telegraph Systems, Conference Paper No. 56-844, AIEE, San Francisco, June, 1956.
[2] MacAdam, W. K., The New Emphasis in Maintenance for Communications Systems, *Trans. AIEE*, **75**, Paper No. CP56-1005, 1956.
[3] McRae, J. W., R. F. Mallina, W. P. Mason, T. F. Osmer, and R. H. Van Horn, Solderless Wrapped Connections, *Bell System Tech. J.*, **32**, 523, 1953. Kuhn, S., Solderless Wrapped Connections, *Bell Lab. Record*, **32**, 41, 1954. Fisher, W. H., Design Trends and Concepts as Applied to Private Wire Telegraph Engineering, *WUTR*, **10**, October, 1956.
[4] Mample, A. Z., Soldered Connections in Electrical Circuits, *WUTR*, **9**, 125, 1955.

and as many pressure contacts, all of which must function perfectly over long periods of time if trouble-free operation is to be achieved. Defective soldering is a conspicuous offender, particularly evident in new work before the "dry joints" have been discovered by inspection, probing, and shaking. A European method[1] puts suspected joints into tuned mechanical vibration while an amplifying a-f circuit picks up detected side bands denoting variable joint resistance while the equipment is under electrical test.

Other continuity defects arise from defective spot welding in tubes and components; "dry" or "unwetted" rivets; dry contacts; loose or oxidized screw-pressure connections; oxidation or poor contact of tube pins and sockets, plugs and jack sleeves, plugs and springs, mountings and fuses or heat coils, crystal holders, rectifier elements; slide and wire contacts; and broken wire or strands inside insulation. Low-voltage a-c measurement of a good soldered connection is of the order of 0.001 ohm; a "dry" connection may measure 0.10 ohm and up.

FIG. 11. Protector for typical telegraph carrier pair.

22. Crosses and Grounds. Shunting of aerial conductors occurs as the result of improper mechanical tension or sag and consequent contacts, or of introduction of conducting material into the lines. When a third, or "foreign," conductor is in contact, it introduces extraneous emf or drains off current, in either case setting up unwanted wave fronts and current flows. Maintenance gangs working on line repairs employ methods to minimize such *swings*, *hits*,[2] *scrapes*, and *crosses*, as they are called—all sources of fortuitous distortion if they are momentary, of low level, and of so infrequent occurrence as not to put a stop to use of the circuit.

Even on insulated pairs, grounding, either direct by wire contact or through lowered resistance introduced by broken insulators, may seriously disturb the coupling of such pairs to others; on grounded circuits it bypasses transmitted energy normally flowing beyond the point of fault.

Akin to a line fault but perhaps more accurately classed as fortuitous distortion are two recognized effects of vibrating line wires: magnetostriction[3] and varying reactance due to changes[4] in L and C between wires.

23. Lightning Interference. Circuit outages due to open and grounded lines, though actually fortuitous, are not usefully classed as examples of fortuitous distortion because of their masking effect. They do affect measurement of reliability of service if not its essential accuracy. In the same class are lightning, aurora, and foreign

[1] Anon., "Maintenance," CCIF Green Book, **3**, 18, XVIIth Plenary Assembly, Geneva, 1954.
[2] Marshall, T. A., Locating Hits on Telegraph Circuits, *Bell Lab. Record*, **19**, 245, 1941.
[3] Dow, M. T., Magnetostriction Noise from Telephone Wires, *Bell Lab. Record*, **22**, 421, 1944.
[4] Packard, P. T., A Line-wire Vibration Damper, *Bell Lab. Record*, **34**, 173, 1956.

voltages, against which the telegraph plant is partially protected at office distributing frames, at junctions of open wire and cable, and, in aggravated cases, at intermediate cable points.[1]

Figure 11 illustrates the *protector*[2] provided at the office distributing frame for a typical carrier-telegraph-line pair. The instrument fuses are normally of 0.8-amp capacity and are designed to protect the office equipment. In the event of contact with a high-voltage power line[3] which would send a large current through a fused arrester to ground, the line fuses are designed to rupture upon passage of a current of 15 amp for 5 min. At open wire-cable junctions, only high-voltage arresters, 750 to 1,200 volts, are used.

Carrier telegraph systems, operating on balanced metallic pairs, are subject to lightning interference[4] because of the low level of the line currents. Much of this interference is caused by the nonsimultaneous operation of the two line arresters because of minor dissimilarities of the arresters or by imperfect balance of the line pair.[5] By including in series with the pair of arresters of Fig. 11 a coil having two closely coupled and carefully matched windings to form an autotransformer with a grounded center tap, it is possible to enforce simultaneous operation of the two arresters. At the same time the presence of the coil in the lightning-discharge path impairs to some degree the protection afforded by the arresters, so that it is desirable to add a second pair of arresters of higher breakdown known as *guard arresters*. Typical breakdown value for the coil arresters is 350 volts and for the guard arresters 750 to 1,200 volts. The combination is effective in the reduction of lightning hits having a peak value sufficient to operate the coil arresters but insufficient to operate the guard arresters.

Where the physical line wires carry no d-c telegraph services, the coil arresters may be omitted so that the coil affords a permanent drainage path for accumulated static voltages as well as lightning hits. In this case,[6] a reduction up to 80 or 90 per cent of the potential atmospheric disturbance can be achieved.

24. Power Interruptions. As serious as an interruption of a telegraph line is a temporary (even momentary) loss of power supply.[7] Wherever possible two sources of d-c supply are tapped. There is an emergency power plant whose prime mover automatically starts and gets under load when the usual d-c voltage drops below a predetermined amount.[8] A storage battery, normally floating, takes the load over during the 0.05-sec switching interval. In the case of radiotelegraph-beam operation[9] where emergency a-c supply is required, plate-supply capacitors hold the voltage during the 0.1 sec required for emergency a-c vibrator units to take up the load. They hold it for ½ min, by the end of which time the emergency alternator is up to speed and automatically takes over.

25. Pole Lines and Aerial Cable.[10] The integrity of open-wire lines begins with the poles,[11] crossarms,[12] and insulators,[13] which support them. Poles are chiefly of yellow

[1] Courtesy of J. R. Hyneman, Western Union Telegraph Company.

[2] Bullard, W. R., J. B. Hays, and H. O. Saunders, Electrical Protection of Telephone Systems, *Trans. AIEE*, **73**, Part 1, 385, 1954.

[3] Bullard, W. R., H. E. Weppler, E. G. Albrecht, H. M. Ellis, *et al.*, Co-ordinated Protection for Open-wire Joint Use (communication—power symposium), *Trans. AIEE*, **75**, Part 1, 215, 1956.

[4] Trueblood, H. M., and E. D. Sunde, Lightning Current Observations in Buried Cable, *Bell System Tech. J.*, **28**, 278, 1949.

[5] Cash, C. C., Protection against Lightning Interference, *Bell Lab. Record*, **15**, 125, 1936.

[6] Duncan, J. A., R. D. Parker, and R. E. Pierce, Telegraphy in the Bell System, *Trans. AIEE*, **63**, 1032, 1944.

[7] Bartlett, I. T., Jr., and H. M. Ward, Mobile Emergency Power, *WUTR*, **7**, 23, 1953.

[8] Steinmetz, A. A., A Precision Voltage Stabilizer for D-c Power Supply, *WUTR*, **3**, 143, 1949.

[9] Ward, H. M., Power Supplies for Microwave Relay Systems, *WUTR*, **3**, 133, 1949.

[10] Courtesy of H. H. Wheeler, Western Union Telegraph Co.

[11] Wheeler, H. H., Engineering of Pole Lines, *WUTR*, **9**, 24, 1955.

[12] Eggleston, R. C., Conversion of Concentrated Loads on Wood Crossarms to Loads Distributed at Each Pin Position, *Bell System Tech. J.*, **29**, 105, 1950.

[13] Wheeler, H. H., and W. F. Markley, Rubber Insulators for Pole Lines, *WUTR*, **3**, 77, 1949.

pine or western cedar. Seven classes have been nationally standardized[1] with dimensions adjusted to give equal strengths regardless of kind of timber. The size and spacing of poles depend upon the expectancy and amount of ice that may form on the wires, the severity of the climate, and the number of wires carried. Based on tests and experience, standard ice and wind loadings for various sections of the United States have been adopted.[2] It is customary in designing lines to provide strengths for the supporting structures that are 2.75 to 3.5 times the theoretical computed values to compensate for the variability and deterioration of poles and to take care of unusually severe storms.

Pine poles are pressure-treated their full length with creosote or pentachlorophenol.[3] Other timbers may have only the lower portion, which rests in the ground, treated.[4] Properly treated poles have a normal life expectancy of 30 to 40 years. At corners, terminals, crossings over railroad tracks, and at the ends of long spans, it is necessary to reinforce the poles by means of guys.[5]

Open-line wires are normally copper, of diameter from 102 to 128 mils, although copper-steel, bronze, or steel conductors of similar diameters may be used for additional strength in longer spans. For 100- to 200-ft spans, which are normally used in communication lines, the *parabolic* formula is used in computing sags and tensions, in place of the more complicated *catenary* formula.[6] Common practice for effecting transpositions, particularly at the higher frequencies, is to employ point-type transposition brackets mounted on crossarms, or in long spans to use span brackets supported by the wires themselves (Sec. 17).

In urban localities, or where the number of circuits required is too large to make open-wire construction practicable, multipair cables, either paper-insulated and lead-sheathed or plastic-insulated and sheathed, are installed either aerially on poles[7] or, where congested conditions aboveground prohibit the use of aerial plants, underground.

Conductor sizes in such telegraph cables vary from 16 to 22 AWG. The number of No. 22 conductors may exceed 1,600 in a cable that is 2⅝ in. in over-all diameter.

Transmission losses in cables vary inversely with conductor diameter. Since even 16-gage conductors have considerably higher losses than open wires, inductive loading must often be resorted to and repeaters must be spaced closer than with open-wire circuits. At junctions of open wire and cable, it is necessary to provide impedance matching for carrier.

Lead-sheathed paper-insulated cables are terminated by splicing them either to the stub cables of gastight sealed terminals or to lead-sheathed rubber or textile-insulated tipping cables, the individual conductors of which are terminated on connecting blocks. Where tipping cables are used, a dam consisting of a suitable wax or resin[8] is provided in the paper cable to make it gastight and thereby prevent moisture from being drawn in by breathing action of the cable due to changes in temperature. Plastic-insulated cables require neither gastight terminals nor tipping cables, since their insulation is relatively impervious to moisture.

Nitrogen introduced into lead-sheathed cables at pressure from 12 to 15 psi is used to locate sheath defects by either pressure gradients, differential-pressure changes, or changes in direction of gas flow. After sheath faults are localized definite locations

[1] "Discussion of National Electrical Safety Code," National Bureau of Standards, Handbook H39, Appendix 3.

[2] "Safety Rules for the Maintenance of Electrical Supply and Communication Lines," National Bureau of Standards, Handbook H32, Sec. 25.

[3] Haenseler, H. A., Wood Preservation, *WUTR*, **7**, 130, 1953. Colley, R. H., The Evaluation of Wood Preservatives (bibliog.), *Bell System Tech. J.*, **32**, 120, 1953.

[4] Leutritz, J., Jr., A Wood Soil Contact Culture Technique for Study of Wood Preservation, *Bell System Tech. J.*, **25**, 102, 1946.

[5] Wheeler, *op. cit.*

[6] Carr, J. A., Tensioning Open Wire for J-Carrier Systems, *Bell Lab. Record*, **23**, 8, 1945.

[7] Gerry, P. R., Aerial Cables and Their Installation, *P. O. Elec. Eng. J.*, **42**, Part 2, 90, 1949.

[8] "Gas Pressure Testing of Cables," Specification 1-A-83, Association of American Railroads, Communication Section, 1942.

are determined with a soap solution or, in the case of major defects, by the sound of escaping gas. The cable is sectionalized by dams inserted at intervals of from 1 to 2 miles before introducing the gas. Important cables are maintained under gas pressure and provided with low-pressure alarms. Alarm gages have a high and low setting; the length of time between the first alarm and the second indicates the magnitude of the leak.[1]

Telegraph cable installed on pole lines is supported by high-strength galvanized or stainless-steel strand, bolted to the poles. The cable is attached to the supporting strand by metal rings or hangers or by corrosion-resistant wires spun spirally about both strand and cable.

26. Underground Cable and Duct. Where cables are placed underground they are either pulled into ducts of vitrified clay, asbestos cement, or fiber or buried directly in the ground. For direct burial, lead-sheathed cables are usually protected with jute and steel-tape armor, while plastic cables are provided with an additional plastic sheath. Manholes are provided at intervals of several hundred feet in duct runs to facilitate installation, testing, and maintenance of the cable.

Lead-sheathed cables installed underground are subject to electrolysis and other forms of electrochemical action due to stray currents and varying earth potentials along the run; so surveys and frequently drainage, cathodic protective systems, or other mitigative measures are necessary to prevent the lead from being destroyed.[2]

TELEGRAPH SYSTEMS

27. Information Theory. Under this title (*cf.* Chap. 1, Sec. 46) are grouped the principles relating to the efficient and accurate transfer of intelligence, discrete bit by bit, from mind to mind. Principles concerning so much of that transfer as may be accomplished by electromechanical means from place to place are sometimes given the narrower title of *communication theory.* The word "modern" is sometimes used to underline the fact that the generalizations and formulas used are recent, though efficiency of codification and the elimination of transmission redundancies are old to the narrow-band telegraph art. Morse code (in its allotment of short, dot-dash combinations) and the American version of Baudot code (in assignment of its one- and two-pulse characters) both bear internal evidence of studied correlation of the briefest code symbols with the highest probability of letter occurrences. Optimum correlations may now be formulated by means of the theories of probability and statistics[3] based upon analogies in thermodynamics[4] and leading to the formulation of ideas of recognition of and choice among messages (aggregations of informational bits) against a background of interference, or noise.[5]

These precepts of communication theory have directed attention to the avoidance of *on-line* transmission redundancies by the expedient of incorporating (*cf.* Sec. 28) qualities of *memory, anticipation,* and *prediction* into receiving equipment.[6] Even more significantly, basic relationships have been demonstrated[7] to exist among signal and noise power, band width, and capacity of a circuit to handle information, which, by raising questions as to ways in which each of these factors may be manipulated for

[1] "Gas Pressure Testing of Cables," Specification 1-A-83, Association of American Railroads, Communication Section, 1942.

[2] Report of Correlating Committee on Cathodic Protection, National Association of Corrosion Engineers, July, 1951.

[3] Darrow, K. K., Memorial to the Classical Statistics, *Bell System Tech. J.*, **21**, 108, 1942.

[4] Shannon, C. E., Prediction and Entropy of Printed English, *Bell System Tech. J.*, **30**, 50, 1951.

[5] Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, *Bell System Tech. J.*, **27**, 109, 1948. Shannon, C. E., Communication in the Presence of Noise, *Proc. IRE*, **37**, 10, 1949. Rice, S. O., Communication in the Presence of Noise—Probability of Error, *Bell System Tech. J.*, **29**, 60, 1950. Schwartz, L. S., Principles of Noise Reduction in Communication Channels, *Trans. AIEE*, **75**, Part 1, 44, 1956. Rice, S. O., A First Look at Random Noise, *Trans. AIEE*, **75**, Part 1, 128, 1956.

[6] Harrison, C. W., Experiments with Linear Prediction in Television, *Bell System Tech. J.*, **31**, 764, 1952. Kretzmer, E. R., Redundancy in Television, *Bell Lab. Record*, **32**, 401, 1954.

[7] Shannon, C. E., A Mathematical Theory of Communication, *Bell System Tech. J.*, **27**, 379, 1948.

the benefit of the others,[1] may influence future telegraph-system designs. The discovered relationship is that the upper limit of the amount of information C measured in binary units (bits) per second, which ideally can be transmitted in one direction over a transmission system in the presence of Gaussian white noise of average power N, is related to the average received power P in the signal and to the band width W occupied by the system expressed in cycles per second according to the equation

$$\frac{C}{W} = \log_2\left(\frac{P+N}{N}\right)$$

Among variants[2] of this relationship are

$$C_1 < W\log_2\left(1+\frac{P}{N}\right) \qquad \frac{C}{W} = 3.3\log_{10}\left(1+\frac{P}{N}\right)$$

$$C = W\log_2\left(k^2\frac{P+N}{N}\right) \qquad C = \frac{H}{T} \leqq 2W\log_2\left(1+\frac{S}{N_v}\right)$$

in which C_1 is *practical* informational capacity rather than ideal; $k \leqq 1$ is a constant defining a desired *quality* of transmission; H is the number of bits in a message and T the number of seconds required to transmit it; S and N_v are, respectively, the *rms* voltages of *maximum* signal and average noise.

Fig. 12. Signals in two-level code A at frequency f and power P are replaceable by equivalent four-level code B at frequency $f/2$ and power $4P$. The intelligence communicated C is the same: bandwidth W has been narrowed by increasing power P, noise N remaining constant. (*From R. M. Page, Comparative Study of Modulation Methods, Trans. IRE*, **CS-1,** *p.* 13, *July*, 1953.)

The basic formula is useful in comparing a given communication system with the ideal and in comparing two systems whose parameters are not greatly different. It will be noted that a low N/P ratio contributes greatly, as a design objective, to the traffic capacity of a telegraph system which has to operate within a limited band width. Conversely, in other instances it may be advantageous to employ wider bands to reduce power.

Each of the factors in the equation is subject to searching reexamination in the light of their relationship. Thus P is not to be considered merely as total power but as composed of *discrete quantized levels* available for the handling of new and more efficient codification of information. From Fig. 12 it will be evident that a direct trade may be made between band width and the number of levels employed in codification.

The end product of communication C is also to be scrutinized in the light of new machinery to handle it—to put it down on paper or to feed it into terminal equipment for the direct operation or control of other machines.

[1] Earp, C. W., Relationship between Rate of Transmission of Information, Frequency, Bandwidth, and Signal-to-noise Ratio, *Elec. Commun.*, **27,** 178, 1948. Clavier, A. G., Evaluation of Transmission Efficiency According to Hartley's Expression of Information Content, *Elec. Commun.*, **25,** 414, 1948. Feldman, C. B., and W. R. Bennett, Band Width and Transmission Performance, *Bell System Tech. J.*, **28,** 490, 1949. Tuller, W. G., Theoretical Limitations on the Rate of Transmission of Information, *Proc. IRE*, **37,** 468, 1949. De Ferranti, B. Z., Information Theory—A Practical Review (British), *Communications and Electronics*, **2,** 52, 1955. Rous, R. F., and R. F. B. Speed, The Significance of Information Theory to Communication Systems (British), *Communications and Electronics*, **2,** 56, 1955. Filipowsky, R., Electrical Pulse Communication Systems (bibliog.), *Brit. IRE J.*, **15,** 451, 1955; **16,** 39, 1956.

[2] Page, R. M., Comparative Study of Modulation Methods, *Trans. IRE*, **CS-1** (1), 13, July, 1953. Anon. (Netherlands), CCIR Document 348-E, VIII Plenary, Warsaw, 1956.

Band width W comes under scrutiny for its present stratification by frequency guard-band losses, and questions are raised as to what extent spectrum space can be conserved by more general use of time-division multiplexes, of types freed by electronics from mechanical limitations.

Especially is C to be examined from the viewpoint of transmission redundancies, raising questions of how much detailed information must be transmitted over circuits and how much can be stored in receivers for use at the behest of control information signals received.

The English language itself, containing countless redundancies and statistical relationships of recurrence frequencies of single letters, diphthongs, digraphs, trigraphs, words, and word sequences, well known to cryptologists,[1] is subject to reexamination in the light of the task of teleprinters to reproduce it. Instead of dealing with one letter at a time, telegraphy may concern itself in the future with the historicity of the pulse sequences transmitted, to ascertain whether they contain statistical properties of movement that can be built into transmitters and receivers, alike, so that only a residue beyond the momentum of each sequence occupies band width in transmission.[2]

By use of storing and searching techniques now commonly built into computers, advantage may be taken of what telegraph transmitters and printers can be made to learn and remember about the sequence of letters in any language. From this advanced viewpoint, a uniform five-unit mark-space code now used in teleprinters is unnecessarily long in that, after each letter is printed, it insists on reserving needlessly a full measure of spectrum for the reception of the next letter. This is not only because the infrequent letter Z and the frequent letter E each contain five pulses, but because, having transmitted in succession a *word space* and the letter Q, present systems assign a full five units (instead of, ideally, one unit or *no* unit) to the letter U which is almost certain to follow.

Suggestions have therefore been made[3] that telegraphy, like television, experiment in new ways to codify intelligence: dealing perhaps not in single letters but in digraph and trigraph sequences; reserving the longest code groups for the greatest departures from normality, and very short groups for normal sequences; telescoping the information at the transmitter; signaling the departures from normal, leaving it up to the receiver to fill in the normals; at the receiver expanding information while decoding it, thus saving band width on the line.

The new approaches therefore trade simplicity of terminal equipment for transmission-spectrum savings. Economics and scarcities are seen to be fundamentals governing research in this field.

The urge to conserve band width by code economics or otherwise varies widely with types of circuit. For local tie lines, where conductor pairs cost as much for teleprinter speeds of 23 cps as for full v-f bands, there is little incentive to compress more information into fewer bits—in fact it has been found advantageous to adopt facsimile, occupying the full band, for certain collateral advantages over teleprinters. But on intercity and international telegraph circuits economics bars the widespread use of wider band widths than are essential to convey information, and here the utility of an extremely short code (approaching two pulses per letter with two-level power[4]) shares interest with more efficient employment of spectrum by multiplexing.

28. Signal Levels and Codes. Facsimile telegraphy, telephoto transmission, and television utilize smooth transitions within a spectrum of gray values between black and white. It has been shown that 128 quantized signal levels may be controlled by a seven-unit code using pulse-position modulation,[5] wherever signal-noise ratios are suitable.

[1] Shannon, C. E., Communication Theory of Secrecy Systems, *Bell System Tech. J.*, **28,** 656, 1949.
[2] Coggeshall, I. S., The Transmission of Intelligence in Typescript, *Trans. IRE*, CS-1 (1), 4, July, 1953.
[3] Oliver, B. M., Efficient Coding, *Bell System Tech. J.*, **31,** 724, 1952.
[4] Shannon, C. E., Recent Developments in Communication Theory, *Electronics*, **23,** 80, April, 1950.
[5] Black, H. S., and J. O. Edson, Pulse Code Modulation, *Trans. AIEE*, **66,** 895, 1947.

For reasons summed up in the description of distortion (Sec. 18) and epitomized by remarks under teleprinter operating range (Sec. 19), most land-line telegraphy is conducted today on a two-level, polar, or f-m basis. Three levels are successfully used in cable Morse where a low level of interference does not preclude one level's being at earth potential. Four levels were, in effect, employed for many years in the land-line quadruplex, wherein a reversal of polarity and a current increment were used for independent signaling.[1]

The 6-*level Squier-Bruce* code of Fig. 13 deserves mention for its beguiling resemblance to a.c.[2] It is an adaptation of Morse, using an arbitrary reversal of polarity

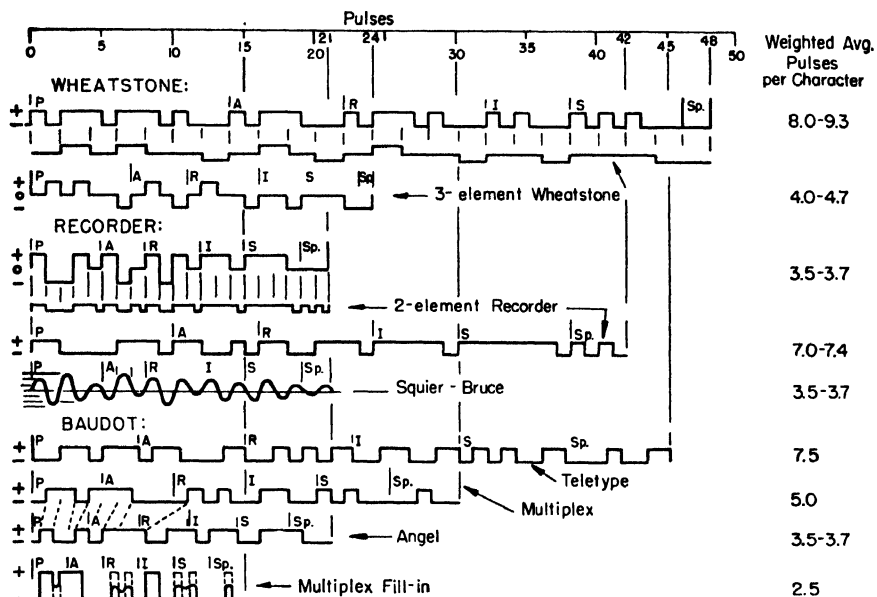



FIG. 13. Comparison of lengths of codes. The first figure is for the "average word" *PARIS* followed by a word-space; the second is for average telegrams in English, including preambles (headings). The codes are described in text. (See also Coggeshall, I. S., *Trans. IRE*, **CS-1**, 4, July, 1953, or *WUTR*, **7**, 56, 1953.)

between successive signals on three levels, dashes taking the maximum applied sine-wave emf, word spaces the minimum, and dots the median level. One of the shortest codes, its wave form resembles a carrier whose frequency exactly coincides with the frequency of intelligence modulation. Its successful transmission would necessitate solution of power-noise relationships at all rms levels, overlapping of side bands, and the suppression of transients inherent in the permutation of signals among six potential levels; the problems resemble those encountered in diplexing f.m. on a.m.

The *Cooke three-unit* code called for three-level transmission, yielding only $3^3 = 27$ combinations, three of which had to be allotted to word-space and carriage shifting, forcing two little-used letters into the upper case.[3] A *four-unit three-level* code would yield 81 combinations—a practical challenge to designers wishing to break with mark-space tradition by adding the third potential level.

[1] Devaux, L., and F. Smets, L. M. T. Laboratories 7-Frequency Radio Printer, *Elec. Commun.*, **17**, 22, 1938.

[2] Squier, G. O., Method of Transmitting the Telegraph Alphabet Applicable for Radio, Land Lines, and Submarine Cables, *J. Franklin Inst.*, **195**, 633, 1923. Squier, G. O., The Application of Radio Engineering Principles to Submarine Telegraph Cables, *J. Franklin Inst.*, **198**, 29, 1924.

[3] Coggeshall, I. S., Submarine Telegraphy in the Post-war [World War I] Decade, *Trans. AIEE*, **49**, 476, 1930.

Figure 13 shows how the British cable system adapts *two-level Wheatstone (Boehme)* Morse for three-level transmission, with a halving of bits per letter.[1] The system is based on the observation that machine-perforated Morse tape, unlike Baudot, contains only three pairs of binaries: M-M, S-S, and M-S (S-M being lacking). The conversion of code levels is accomplished at repeaters where cables (safe for three-level transmission) and land lines (with their distortion problems) meet. Figure 13 also

Table 4. Design Constants—Five-unit Start-Stop Teleprinters

5-unit intelligence; No. of unit pulses in transmission pattern	Nominal[a] transmitting speed			Nominal[b] pulse length				Millisec per character	Where used
				In units		In millisec			
	Operations per min	Avg[c] words per min	Bauds	Start and code pulses	Rest, or stop, pulse	Start and code pulses	Rest, or stop, pulse		
7	390[d]	65	45.5	1	1	22	22	154	Western Union teleprinters, U.S.
7.42	368[d]	61.3[f]	45.5	1	1.42	22	31.24	163.2	Bell System, U.S.[e]
7.42	460	76.67[h]	55.65	1	1.42	17.58	25	130.48	U.S.[i]
7.42	600	100	74.2	1	1.42	13.48	19.12	100.0	U.S.[i]
7.5	400[g]	66.67	50	1	1.5	20	30	150.0	CCIT standard, Europe
7	428.6[g]	71.4	50	1	1	20	20	140	Former European standard
7.42	404[g]	67.3	50	1	1.42	20	28.4	148.4	U.S. military, for interoperation with allied military

Courtesy of F. W. Smith, Western Union Telegraph Company.

[a] The nominal transmitting speed may vary by not more than 1 per cent because of compromises in gear ratios made by various manufacturers.

[b] The nominal pulse lengths may vary by not more than 1 per cent because of compromises in gear ratios and rounding off of angles between transmitting cams or the included angles of faceplate segments. Actual pulse lengths may vary by as much as 3 per cent when signals are generated by a faceplate type of distributor or distributor-transmitter, and by as much as 5 per cent when signals are generated by cam-operated contacts, as on a keyboard transmitter.

[c] An "average" word is considered to be six characters long, including the space following the word and any function characters transmitted, such as letters shift or figures shift.

[d] These two codes are compatible.

[e] 7.42-unit equivalent frequently referred to informally as "seven and a half' unit codes.

[f] 368 o.p.m. is frequently referred to informally as "60" words per minute, or "60-speed."

[g] These three codes are compatible.

[h] 460 o.p.m. is frequently referred to informally as "75" words per minute.

[i] 7.42 equivalent employed by all United States users at 460 and 600 o.p.m.

shows how the reverse takes place in the use of the *Higgitt* code in lengthening cable-recorder Morse to two levels when passing from cables to radio or land lines.

The uniform-length *two-level five-unit teleprinter* code[2] (see Chap. 24, Secs. 3 to 13) used in radiotelegraphy is in practically universal use in wire operations. Figure 13 affords a comparison of its length—in its multiplex, start-stop, and submarine-cable versions—with other wire-telegraph codes. Only in its time-division-multiplex application (Sec. 36) does it perform its whole function with five transmitted units; its

[1] Coggeshall, I. S., The Transmission of Intelligence in Typescript (bibliog.), *Trans. IRE*, **CS-1** (1), 4, July, 1953.

[2] Harris, R. W., Keyboard Standardization, *WUTR*, **10**, 37, 1956. Gilbert, E. N., A Comparison of Signalling Alphabets, *Bell System Tech. J.*, **31**, 504, 1952. Hayton, T., C. J. Hughes, and R. L. Saunders, Telegraph Codes and Code Convertors, *Proc. IEE* (London), **101**, Part 3, 137, 1954. Earp, C. W., Recent Development in Communication Technique, *Elec. Commun.*, **30**, 61, 1953. Buff, C., Twinplex and Twinmode Radiotelegraph Systems, *Elec. Commun.*, **29**, 20, 1952.

employment with addenda of start and rest pulses of various lengths is shown in Table 4, together with data relating to the printers using five-unit code (Baudot). Its $2^5 = 32$ possible combinations are distributed among 26 alphabetic letters and 6 printer manipulating functions, chief among which are the two which change the printer case to introduce 10 numerals, punctuation, and additional control signals. Two of the latter controls could, of course, put a modified printer into third case with the same code, and such machines have been built experimentally. An analogous more practical extension of the five-unit code's capabilities is employed in the *stunt* mechanisms of modern printers.

Six-unit general-purpose printers have been proposed to afford $2^6 = 64$ combinations, making possible easy adaptation of printers to the typewriter's lower-case type font, or giving scope to language alphabets containing more than 26 characters as commonly printed. *Teletypesetters*[1] fall in the six-unit category. A growing number of accounting,[2] inventory,[3] computing,[4] and other business machines require more than five units (Sec. 37).

Stock and commodity printers use a six-unit modification of five-unit Baudot code, adopted because of the occurrence of many shifts to upper case in ticker tape. The presence or absence of the sixth unit of the code determines the case in which the Baudot-selected character will print.

By allotting to the most frequently used letters in English the minimum number of marking pulses, Baudot code saves wear and tear on machines and reduces the average number of polar signal transitions on the line per character handled on start-stop machines to two,[5] thus aiding accuracy. *Angel's* shortened version[6] of Baudot was successfully used on submarine cables. The marking bits of Baudot were transmitted at ordinary length and frequency, but there was an arbitrary reversal of polarity between successive marking pulses. The spacing impulses were transmitted half length each, any number of them in succession being joined as to polarity with the marking impulse which preceded them. The characteristic uniform length of Baudot characters was sacrificed. Angel's choice of *mark*, rather than *space*, for full-length transmission conformed with the Baudot code's correlation of the fewest mark bits per character with the most frequently used characters.

FIG. 14. Gulstadt fill-in circuit. Original transmission as in *C*. Nearly attenuated single impulses and fully transmitted doubles (and more), after transmitting line, shown in *A*. In *B*, singles have been fully suppressed by amplifier shaping; in *C* they have been locally filled in, thus wholly restored.

Modern ocean-cable transmission utilizes a *Gulstadt vibrating fill-in* circuit[7] which multiplies by a factor of 1.8 the traffic-carrying capacity of unmodified Baudot. Its effect in contracting code length is depicted on the last line of Fig. 13. The system calls for the transmission of ordinary five-unit tape at a rate so high that single bits are almost attenuated,[8] as shown in Fig. 14*A*. The noise level in the cable associated with single bits becomes no longer a factor in their reception; the materiality of noise N may thereafter be related to the signal power P included in lobes composed of *two*

[1] Smalley, M. N., Teletypesetter Equipment in the Bell System, *Bell Lab. Record*, **31**, 183, 1953.
[2] Ostline, J. E., Automatic Call Recording and Accounting, *Communications and Electronics*, **1**, 56, 1953.
[3] Schmidt, E. L., and J. J. Connolly, Inventory System Using Digital Computer Techniques, *WUTR*, **5**, 148, 1951.
[4] Richards, R. K., "Arithmetic Operations in Digital Computers," Van Nostrand, 1955.
[5] Cory, S. I., Telegraph Transmission Coefficients, *Bell Lab. Record*, **33**, 11, 1955.
[6] Angel, H., Printing Telegraphs on Non-loaded Ocean Cables, *Trans. AIEE*, **46**, 884, 1927.
[7] Haglund, H. H., and A. W. Breyfogel, Thyratrons in Loaded Cable Multiplex, *WUTR*, **3**, 34, 1949.
[8] Finnegan, E. T., Magnetic Tape Signal Transmitter for Ocean Cables, *WUTR*, **10**, 112, 1956.

bits, to which the receiver circuits are resonated with the effect shown on line *B*. The attenuated single bits are then interpolated locally by flat-top a.c. generated by an auxiliary ring on the distributor, restoring, as shown on line *C*, the original shape of the signal train including all the single bits. The system takes advantage of the filtering action of a long cable, worked to the maximum traffic capacity its band width will permit, in suppressing information which the receiver can predict—in this instance that, upon decay of a bit or a succession of like bits, the next bit will be of opposite polarity whether or not strong enough to be sensed by the receiver.

29. Wire and Radio Systems. Morse and teleprinter codes, described in Chap. 24 as used on radiotelegraph circuits, are used on wire-telegraph systems. The importance of Morse has sharply declined in American land-line and submarine-cable practice. Its present chief uses are in connection with regulating American circuits in cases where the telephone is not used, in the world-wide British cable system, and on various European lines.

Teleprinters or teletypewriters, under the trade names of Teletype, Creed, Siemens & Halske, Olivetti, and others, are in almost universal use in the United States and are widely used in Europe. The majority of them operate on start-stop synchronism as typified by the type 15 Teletype and the compatible[1] Western Union start-stop teleprinter, the differences between which are shown in Table 4. Distributor-transmitter adjuncts,[2] operating "off-line" from prepared five-unit tape instead of directly from the "on-line" keyboard, are in general use wherever it is expedient to reduce to zero the line time lost by slow manipulation or imperfect keyboard cadence, or wherever it is desired to precheck transmission initiation accuracy.

It is the practice to operate teleprinters in country-wide end-to-end circuit connections over telegraph carrier facilities of both the telegraph and telephone companies. Beam, microwave, and scatter-beam radio links are used in certain cases where expedient and economic. Radiotelegraph and submarine-cable links overseas also interconnect American and European teleprinters.[3] Keyboard-typebar-code standards, based upon international alphabet No. 2, apply alike to European and American land-line and overseas practice. To accommodate alphabet differences among the various languages, the standards have to provide for certain choices, easy to arrive at by agreement between any two countries but presenting more difficulty when two continents are involved. The differences, which are being eliminated or accommodated by automatic *code translators*, are more acute as to "stunt functions" of teleprinters and to punctuation than to alphabets and numerals. The European speed standard (50 bauds; *cf.* Table 4) and its 20-millisec pulse length compared with 45.5 bauds and 22 millisec in the United States also require specialized repeaters to reconcile.

Telegraph circuits may be classified in two ways: those which are more or less permanently set up between terminals (*trunks*) and intermediate (*way*) stations (*drops*), and those which are usually made up of component line sections temporarily joined by switching. Switched circuits may again be classified in two ways: those which are made up of line sections closely coupled by direct connection or repeaters; those whose line sections are more loosely coupled by reperforators (tape storage), magnetic drums (memory devices), or variants. Both permanent and switched circuits may be one-way, two-way for simultaneous operation (duplex), or two-way back-and-forth (half-duplex).

Both the radio and the wire companies lease their facilities to governmental and commercial users for private-wire services. These are almost universally teleprinter circuits, the simpler installations consisting of two printers and a two-way "line," working one way at a time (half-duplex) or both ways simultaneously (duplex).

[1] Smith, F. W., Teleprinter Codes: 7-Unit versus 7.42-Unit, *WUTR*, **8**, 129, 1954.

[2] Harrison, H. H., Principles of Modern Printing Telegraphy, *J. IEE* (London), **54**, 309, 1916. Harrison, H. H., Developments in Machine Telegraph Systems and Methods of Operation, *J. IEE* (London), **68**, 1369, 1930.

[3] Sparks, S., and R. G. Kreer, Tape Relay System for Radio-telegraph Operation, *RCA Rev.*, **8**, 393, 1947. Dietsch, C. G., The Tangier Radio Relay System, *RCA Rev.*, **14**, 557, 1953.

With the multiplication of outlets, private networks are frequently developed, with provision for switching and storage.

The telephone companies in the United States furnish switched telegraph interconnections for TWX subscribers by means of teletypewriter exchange service, employing many of the switching methods and devices applicable to long-distance telephone calls.[1] A similar network of the various telegraph administrations covers Europe under the name "Telex." The American radio companies furnish overseas connections between TWX and Telex.

Public telegram service is furnished abroad by post office authorities who also administer telephone service. In the United States the telegraph company renders the message service,[2] employing its own facilities and those it leases from the telephone companies. The use of telegraph carrier to subdivide owned or leased voice bands is typical for intercity connections. Interconnected major reperforator switching centers are located strategically to serve all other telegraph points in their territories as tributary offices. Reperforator offices are by-passed in certain cases where community of telegraph interest between two cities calls for direct trunk circuits. Microwave radiotelegraph is employed on selected trunk routes of high traffic density justifying the capital expense.

Radiotelegraph companies have established central offices in international gateway cities for handling overseas the same types of service in similar fashion, using similar land-line circuits and equipment.

Problems associated with connections to branch offices, subservient to central telegraph offices in a given large city, and with telegraph connections of the land-line, radio, and submarine-cable common carriers to their customers[3] absorb a great deal of administrative and engineering thought but are considered beyond the scope of this handbook except for the treatment of *loops* and *legs* (Sec. 31). Such terminal handlings are usually accomplished by teleprinter *tie lines*, *facsimile* tie lines, by the public telephone service or special telephone tie lines, by underground pneumatic tubes,[4] or by foot, mounted, or automobile messengers. Of these methods, facsimile transmission is the fastest growing for the purpose stated, teleprinter the one most compatible with intercity transmission and the only one widely employed by the telephone companies.[5]

30. Start-Stop Teleprinter Systems (see Chap. 24, Sec. 18). The simplest system, say for intraplant dispatches, consists of two compatible a-c- or d-c-powered teleprinters in the same locality working together alternately back and forth over a grounded line or metallic loop. The line, supplied with d.c. or rectified a.c., will stand normally closed on the rest (stop) segments in the printers. Speed matching is unnecessary so long as both printers are a-c driven from a 60-cps supply source; d-c machines will require a simple stroboscopic speed adjustment, checked with a portable tuning fork. Operating *margin,* or range, measurement and adjustment are built in. Electrical protection must be provided. Supervisory alarms and printer-control signals operable upon initiation and termination of calls or faulty line conditions are optional. A system of exchange of serial-message-number information guards against line and operator failures.

[1] Duncan, J. A., R. D. Parker, and R. E. Pierce, Telegraphy in the Bell System, *Trans. AIEE*, **63**, 1032, 1944.

[2] D'Humy, F. E., and P. J. Howe, American Telegraphy after 100 Years, *Trans. AIEE*, **63**, 1014, 1944.

[3] Organ, N., *et al.*, Central Office Engineering of Modern Telegraph Offices, *WUTR*, **5**, 92, 1951.

[4] Stieber, W., Automatic Switching in Pneumatic Tube Systems, *Elec. Commun.*, **27**, 260, 1950.

[5] Murray, D., Speeding Up the Telegraphs—A Forecast of the New Telegraphy, *J. IEE*, **63**, 245, 1925. The Development of Western Union [Plan 21-A Selector] Switching System, *WUTR*, **3**, 89, 1949. Agnew, E. J., Automatic Switching of Telegraph Messages, *Elec. Commun.*, **27**, 280, 1950. Watson, E. F., Fundamentals of Teletypewriters Used in the Bell System, *Bell System Tech. J.*, **17**, 620, 1938. Cofer, T. F., (Start-Stop) Printing Telegraphy and Inductive Disturbances, *WUTR*, **1**, 48, 1947. Easterling, C. E., and J. H. Collins, The Teleprinter No. 11 (Creed), *P.O. Elec. Eng. J.*, **46**, Part 2, 53, 1953. Anon., Creed No. 47 Tape Teleprinter, *Elec. Commun.*, **25**, 421, 1948. Easterling, C. E., Automatic Teleprinter Working, *P.O. Elec. Eng. J.*, **47**, Part 2, 80, 1954. Collins, J. H., and E. G. Collier, The Testing of Telegraph Machines, *P.O. Elec. Eng. J.*, **49**, Part 1, 7, 1956.

Apart from make of teleprinters selected, a choice must be made as to whether they are to constitute a *page*-producing or a *tape*-producing system. Printer tape is gummed on the back and affixed with a hand-held moistener. Tape operation facilitates neat correction of most errors by a paste-over technique, with consequent economic advantage. Page printers for use in their own offices have more appeal to business users[1] and are more versatile in such business tasks as preparing carbon copies and forms. Since keyboard and type-pallet associations with the Baudot code differ as between page and tape printers, the two types are incompatible in a system unless end-of-line attachments are placed on keyboards and complex code translators are placed ahead of printers.

Choice of printer speeds will ordinarily fall among those most in use, nominally 60, 65, 75, and 100 wpm. The 75-speed lies on the outer margin of continuous operational ability of a typist capable of producing 100 to 125 wpm in spurts on a typewriter. However, distributor-transmitters may be procured which, even more effectively than any operator, will occupy full line time at 75- and 100-speeds. Choice of speed is an economic balance between traffic-load totals, traffic-peak requirements in each direction, expected average and maximum speed of service required, distance between terminals, line-rental rates, equipment-rental rates, and operators' and maintainers' salaries allocable to the telegraph operations. Marginal situations sometimes control choice of whether to pay for premium speed, adopt full-duplex instead of half-duplex, or add another circuit.

The teleprinter is normally operated as an entity, the machine at each end of the circuit printing, in turn, a home copy of its own sending, then copy received from the distant end. However, the keyboard and typing mechanism may be split, at will, for duplexing the line (two ways simultaneously). In that event the home copy is suppressed, and the operator sends blind; meanwhile the typing mechanism remains responsive to signals from the distant end. If home copy is desired, two machines are used at each end of a duplexed circuit.

For more typical operation between different localities, a user will secure lines from the telephone or telegraph common carriers; both compete for the business and include engineering counsel as part of their service; the same observation would include the radiotelegraph companies for service overseas. Customers' teleprinters and associated tie-line loops, including power-supply devices, are usually recommended by the carrier and furnished on a rental basis; in today's practice they actuate intercity carrier modulators in the companies' central offices.

31. Loops and Tie Lines. Teleprinters, as either sending or receiving instrumentalities, are usually connected with central telegraph offices by subscribers' loops (two conductors in cable) or legs (single conductors). At the central office the loop or leg works through either a repeater or relay or directly to actuate an intercity channel, neutral, polar, or carrier.[2] The availability of power on the subscriber's premises sometimes determines the type of loop or leg operation adopted. A-c lines will supply the teleprinter motor, but if loop or leg arrangements make it necessary to supply neutral or polar d.c., it must usually be obtained through use of rectifiers.

The advantages accruing to polar main-line operation (Sec. 14) carry over to leg operation. A case can be made for universal use of polar legs and loops;[3] a trend in that direction has been facilitated by the availability of cheaper sources of rectified polar potentials and simpler polar relays; and further impetus will be lent by the fact that the new 28-type printer is polar-equipped. For regulating facility it is desirable to have all legs converted to polar operation if any are.[4] At present the 15-type tele-

[1] See Sec. 43 for description of Teletype Model 28. Smith, Fred W., Modern (Kleinschmidt 150) High-speed Page Teleprinters, *WUTR*, **9**, 77, 1955.

[2] Davey, J. R., Circuit Features of the No. 2 Telegraph Serviceboard, *Bell Lab. Record*, **33**, 424, 1955.

[3] Jaeger, E. F., Centralized and Simplified Circuit Handling Using Polar Leg Operation, *WUTR*, **8**, 2, 1954. Shafer, D. P., Comparison of Make-Break and Polar Operation, *WUTR*, **8**, 81, 1954.

[4] Davey, J. R., F. H. Hanley, and M. R. Purvis, A New Telegraph Serviceboard Using Electronic Circuits, *Trans. AIEE*, **74**, Part 1, 30, 1955.

printer requires a supplementary relay for polar operation. Certain carrier channel legs are worked without relays,[1] simple closing and opening of a loop by the subscriber having the effect of using central-office d-c battery to shift the carrier frequency; but polar loop operation may be used if desired (Fig. 17).

In Bell System practice[2] f-m carriers are manipulated over d-c on-off loops in cable for as long as 25 miles without need of correcting any loop-signal distortion.[3]

32. Telegraph Carrier Systems. Typical modern wire-telegraph transmission between cities is accomplished by amplitude[4] or frequency modulation[5] of continuously

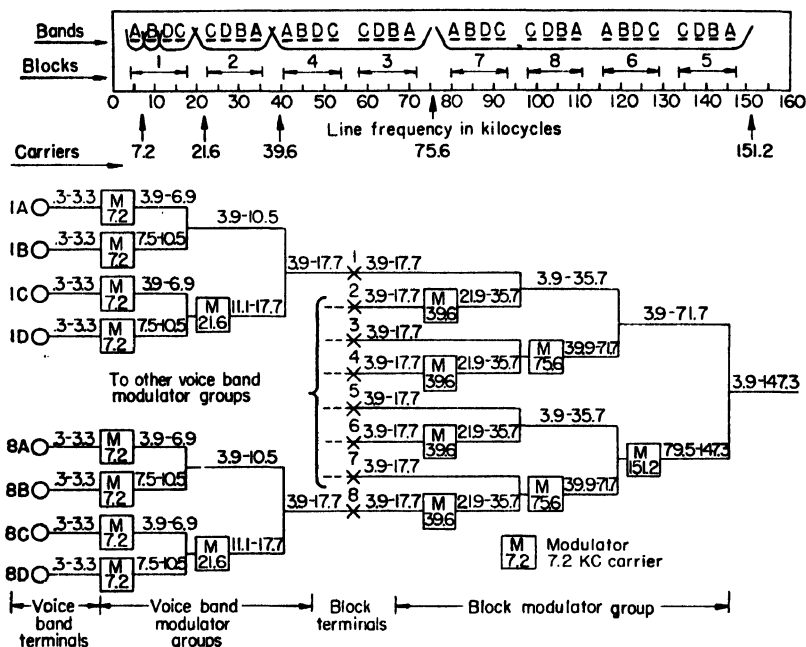



FIG. 15. Frequency allocations in one direction in a 150-kc telegraph system. All v-f channels in all eight blocks are shown in the spectrum layout. The modulator layout shows only the outer blocks, 1 and 8. There are 16 teleprinter channels in each v-f band. The number of v-f translations each undergoes varies from five (in 8C or 8D) down to one (in 1A and 1B); there is an additional translation for channels in each upper subband of v.f.

generated a-f carrier channels, stacked above each other by processes of frequency-division multiplexing and frequency translation to fill the available spectrum of the wire circuit, as shown in Fig. 15.

The 3-kc v-f band is the vehicle of all present-day long-haul telegraphy. When carried along with voice bands in use for telephonic conversations on the telephone companies' wires, the center spacing of v-f bands is 4 kc. Each v-f band is broken down in telegraph-company practice to two subbands and each subband into 9 channels with centers 150 cps apart. In American telephone practice the separation of center frequencies for telegraphy is 170 cps; in British Post Office practice, 120 cps.

[1] Cusack, F. H., and A. E. Michon, An F-m Telegraph Terminal without Relays, *WUTR*, **1,** 33, 1947.

[2] Bogan, L. B., and K. D. Young, Simplified Transmission Engineering in Exchange Cable Plant Design, *Trans. AIEE*, **73,** Part 1, 498, 1954. Purvis, M. R., Service Features of the No. 2 Telegraph Serviceboard, *Bell Lab. Record*, **33,** 100, 1955.

[3] Shanck, R. B., Flexibility of the 43A1 Carrier Telegraph System, *Bell Lab. Record*, **34,** 175, 1956.

[4] Matte, A. L., Advances in Carrier Telegraph Transmission, *Bell System Tech. J.*, **19,** 161, 1940.

[5] Bramhall, F. B., and L. A. Smith, A Nation-wide F-m Telegraph Network, *Trans. AIEE*, **70,** 338, 1951.

The Bell System's a-m system subdivides a v-f band into 20 telegraph channels; its f-m system[1] provides as many as 17 telegraph channels in the v-f range.[2]

It will be appreciated that in carrier telegraphy several scores of telegrams are being handled simultaneously over the same unidirectional pair of conductors, often in association with telephone, telephoto, and other modulations being applied at the same time. Nevertheless, each unit impulse of each letter of each telegram imparts its influence to the wave shape of the composite applied emf; and the line, filters, and other common apparatus respond accordingly. The designer has a choice of whether to employ a-m or f-m type of modulation for signaling on the individual telegraph channels.

FIG. 16. F-m telegraph signals in sending oscillator (D) and at receiver before limiting (G), showing influence of sidebands on receiving amplitude.

By use of f.m., freedom is secured from errors due to changes in transmission equivalent on the vehicle channels and from errors due to extraneous disturbances, equivalent to about 9 db, from all sources;[3] and by taking advantage of f.m.'s equal-valued signal strengths of different frequencies, the telegraph system works, as in polar operation, independently of the magnitude of the incoming envelope of the carrier compared with a local reference level, and free of mark-space bias.

On each a-m frequency-division channel of communication the carrier or a subcarrier has its form altered by the modulation frequency and supplemental power

[1] Davey, J. R., and A. L. Matte, Frequency Shift Telegraphy—Radio and Wire Applications, *Bell System Tech. J.*, **27**, 265, 1948. Matte, A. L., The 40AC1 Carrier Telegraph System, *Bell Lab. Record*, **27**, 327, 1949. Hysko, J. L., W. T. Rea, and L. C. Roberts, A Carrier Telegraph System (43A1) for Short-haul Applications, *Bell System Tech. J.*, **31**, 666, 1952.

[2] Shanck, *op. cit.*

[3] Bramhall, F. B., and J. E. Boughtwood, Frequency-modulated Carrier Telegraph System, *Trans. AIEE*, **61**, 36, 1942. Cusack, F. H., and A. E. Michon, A F-m Telegraph Terminal without Relays, *Trans. AIEE*, **66**, 1165, 1947.

added to it by pairs of modulation side frequencies.[1] Similarly, f.m. influences the form of carrier as shown in Fig. 16. The carried wave C varies the instantaneous frequency of carrier A. Close examination of the line wave D will show employment of unity modulation index, the carried wave being 35 cps for teleprinter operation[2] with 150-cps spacing between adjacent channels, and the line wave being deviated 35 cps either side of its unmodulated frequency. Since the carrier wave in f.m. is not furnished supplemental power by the side bands E and F, the side-band power is contributed by the carrier as shown in G; the second- and third-order side bands which constitute less than 3 per cent of line wave power are eliminated in practice.[3]

The channel transducer which converts transmitting loop d-c signals into corresponding channel frequency modulation is shown in Fig. 17. The rated output power of the oscillator-amplifier is 10 dbm. Transients in the tuned oscillator circuit, due to sudden transition of the telegraph impulse which might persist even longer than the

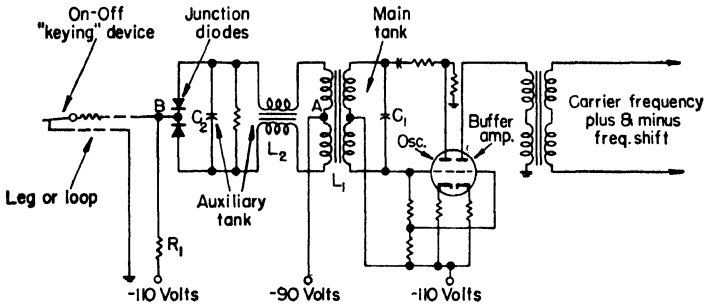



FIG. 17. Modulation of f-m telegraph carrier channel by leg or loop from subscriber. No relays are required.

duration of the transition interval, are overcome by the circuit configuration represented by the differential transformer, the split inductor, and the twin diode. C_1 remains in circuit at all times; C_2 is cut in and out of the tank circuit in response to keying of the loop circuit. With the sending loop open it draws no current and the potential of A is 20 volts positive with respect to B. The diode is conductive and acts to cut C_2 out of the tank circuit. L_1, L_2, C_1 are proportioned to produce the spacing frequency. When the sending loop is closed, a drop of 40 volts in R_1 causes A to become 20 volts negative with respect to B; the diode becomes nonconducting; C_2 is switched into the tank to lower the oscillator frequency. The marking frequency f_m is just 70 cps below f_s. The shift is rather abrupt but as observed at the output of the sending-end channel filter is gradual as dictated by the band width of the filter and the reluctance of its "frequency time constant" to accommodate a shift in frequency. As a result, looked at from the receiving end before limiting, the frequency change is practically sinusoidal, just completing one transit in time to start the next (Fig. 16).[4]

Such a channel, centered, say, on 825 cps and deviated ± 35 cps, will accommodate, on a nationwide basis, 100-speed start-stop channels (75 bps, 37.5 cps) if desired; on a more restricted basis certain of them are carrying three-channel time-division-multi-

[1] Cusack, F. H., Frequency Translation in Wide-band Transmission Systems, *WUTR*, **2**, 36, 1948. Jones, T. A., and K. W. Pfleger, Performance Characteristics of Various Carrier Telegraph Methods, *Bell System Tech. J.*, **25**, 483, 1946.

[2] Boughtwood, J. E., A 150-kilocycle Carrier System for Radio Relay Applications, *WUTR*, **2**, 57, 1948.

[3] Wickizer, G. S., Relative Amplitude of Side Frequencies in On-Off and Frequency-shift Telegraph Keying, *RCA Rev.*, **8**, 158, 1947. Jatlow, J. L., and B. B. Mahler, Carrier Telegraph System Using Frequency Modulation, *Elec. Commun.*, **27**, 205, 1950.

[4] Holland, J. D., Design Considerations for a Radiotelegraph Receiving System, *Elec. Commun.*, **29**, 34, 1952. Lyons, W., Design Considerations for Frequency-shift-keyed Circuits, *RCA Rev.*, **15**, 230, 1954. Laport, E. A., and K. L. Neumann, A New Low-power Single-sideband Communication System, *RCA Rev.*, **16**, 635, 1955.

plex printer circuits at a total speed of 200 wpm (100 bps, 50 cps). Its ordinary loading consists of a one-way 23-cps teleprinter. Eight companion channels, with eight pairs of generated f-m values, share with it a subband, shown in Fig. 18.

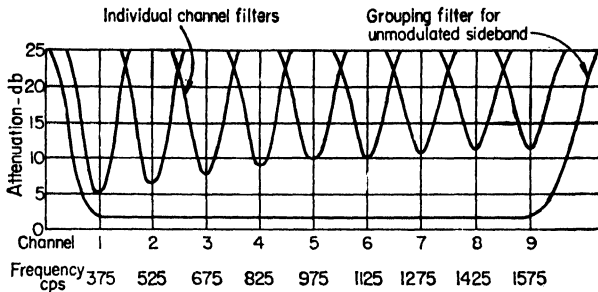



FIG. 18. Channel allocations in primary sub-v-f band. Separation between centers, 150 cps.

A complete small carrier system could be so devised, and, as a matter of fact, in telegraph-company practice, even when larger ones are compounded of stacks of such half v-f bands, only these nine pairs of frequencies are generated by the oscillators. In the telephone company's a-m practice, where 20 telegraph channels occupy a v-f band, the 20 are separately generated.

In either case, the generated groups are stacked, or translated upward into their proper individual occupancies of the upper parts of the transmission spectrum,[1] by double-balanced modulators,[2] and at the receiving terminal inversely demodulated to produce the original basic f-m channel signals for detection, discrimination, and printing.

In Western Union practice,[3] two nine-channel subbands are placed together, by translating one of them, to form a v-f band (Fig. 19). In its use of Bell System carrier facilities, Western Union modulates each v-f band in this way. Alternatively, when it uses its own lines, four v-f bands are stacked to form a block, and two blocks will constitute a 30-kc carrier system. Or four blocks may be stacked to constitute a block modulator group; and two of these groups, comprising 64 subbands, compose the 150-kc SSB suppressed carrier system,[4] with $P/N = 53$ db. In Fig. 20, one such 150-kc system is shown as f-m modulator of 1,000-kc f-m subcarrier of an r-f micro-wave-beam carrier in the 3,900- to 4,200-Mc band, deviated ± 2 Mc.[5]

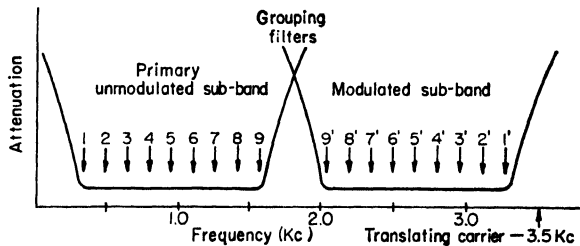



FIG. 19. Translation of v-f subband to fill voice band. Primary modulated subband is that of Fig. 18. Modulation with the translating carrier produces primed telegraph channels in second subband. Entire v-f band becomes building block for further translations of Fig. 15.

[1] Cusack, F. H., Frequency Translation in Wide-band Transmission Systems, *WUTR*, **2**, 36, 1948.
[2] Bramhall, F. B., The Suppressed Carrier Modulator as a Frequency Translator, *WUTR*, **1**, 23, 1947.
[3] Boughtwood, J. E., A 150 KC Carrier System for Radio Relay Applications, *WUTR*, **2**, 57, 1948.
[4] Millar, J. Z., Two Thousand Telegrams a Minute by Microwave, *WUTR*, **1**, 2, 1947. Corwith, H. P., and W. B. Sullinger, Western Union's Microwave Relay System, *WUTR*, **2**, 99, 1948.
[5] Gerlach, G. G., A Microwave Relay Communication System, *RCA Rev.*, **7**, 576, 1946. Hansell, C. W., Development of Radio Relay Systems (bibliog.), *RCA Rev.*, **7**, 367, 1946. Knox, J. B., and C. H. Brereton, A Multi-channel VHF Radio Communications System, *RCA Rev.*, **7**, 179, 1946. Boughtwood, J. E., Telegraph (Radio) Terminal AN/FGC-29 Circuit Design Aspects, *Trans. AIEE*,

33. Switching and Storage Systems. Reperforators—receivers which differ from printers in that they re-create punched tape to be run for onward retransmission rather than page "copy" or gummed typescript tape—have found wide application in connection with the temporary storage of telegraphic information and its switching, either manually or automatically, to other circuits. Typical reperforators (Sec. 44) reproduce at a distance perforated five-unit Baudot tape, although designs using six-unit *uniform-length code* and Wheatstone (Boehme) Morse nonuniform code are in widespread use. The number of unit signals across a tape and the number of center holes (for transmitter-feed purposes) per character may vary as required up to mechanical limitations. Sensing of coded punched holes is usually accomplished by the introduction of transmitter pins which actuate contacts, although photoelectric

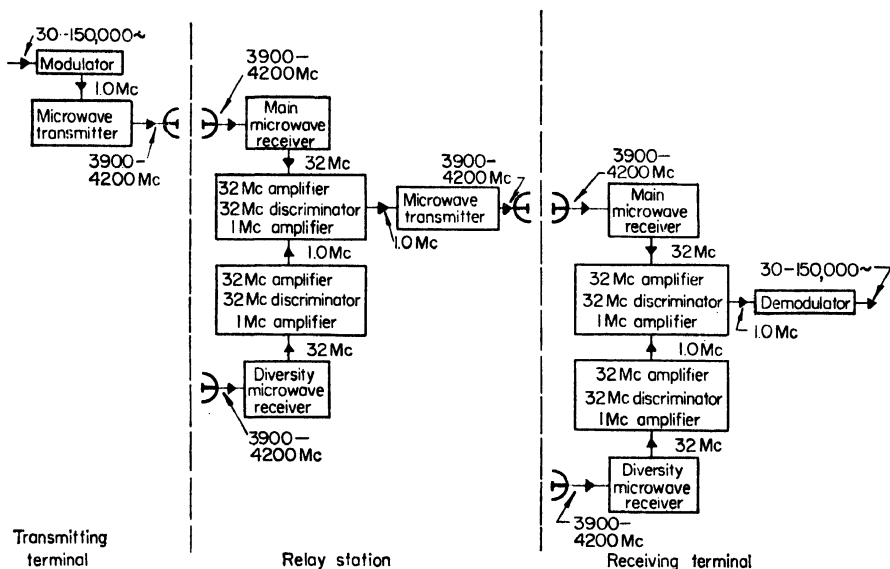



FIG. 20. One 150-kc (32 v-f) band of Fig. 15 applied to microwave beam. Double f-m modulation is employed to translate broad-band telegraph spectrum through 1 Mc to 4,000 Mc at transmitters and through 32 Mc and 1 Mc at receivers. One repeater station is shown. Space-diversity reception is employed.

means afford an alternative. For higher speeds, magnetic tape and drum storage and xerographic[1] processes are being experimentally employed.

Reperforators are ordinarily organized into switching systems in which their characteristic function is the temporary storage of information at junction points in a chain of circuits. Such loose coupling lends itself to one-way operation of telegraphs, with economic advantages flowing from: (1) relief from the necessity of setting up a circuit route all the way from point of origin to destination of a telegram in order to avoid manual retransmissions en route, (2) consequent widespread interchange utility of circuit sections, (3) reduction in holding time in channel-mile-minutes, and (4) reduction of idle time on that half of a two-way circuit held inoperative in back-and-forth working. Reflecting these economies, the reperforator has radically changed the physical

73, Part 1, 531, 1954. Cusack, F. H., Telegraph (Radio) Terminal AN/FGC-29 Equipment Features, *Trans. AIEE*, **73,** Part 1, 536, 1954. Mack, A., and R. H. Levine, A New Multichannel Teletype Terminal for Use on Long-range H-f Radio Systems, *Trans. AIEE*, **73,** Part 1, 540, 1954. Higgins, C. A., and E. A. Gilbert, A New Audio Telegraph Carrier Terminal, *Trans. AIEE*, **74,** Part 1, 58, 1955.

[1] Bramhall, F. B., The Xerographic Process, *WUTR*, **3,** 43, 1949. Young, C. J., and H. G. Greig, Electrofax Direct Electrophotographic Printing on Paper, *RCA Rev.*, **15,** 469, 1954.

aspect of central telegraph offices, land line, cable, and radio alike, and has been the prototype for other information-storage devices, such as those used in computers and card-punch business machines.

The versatility of reperforation, as a system, for serving commercial telegram, private leased telegraph wire, and teletypewriter exchange services has been due in large measure to the highly developed switching techniques[1] of telephony (*cf.* Chap. 28, Secs. 14 to 19). Pulse-controlled and relay-operated ringing, signaling, and supervisory systems by which telephone voice circuits were at first manually established by operators and later set up by dial-operated machine switching constituted the first widespread intercity use of nonlanguage "data," or bit-by-bit information (Sec. 37).

In telegraphy as in telephony, the principal switching components are two-position multicontact relays and rotary and crossbar switches. Transmission accomplishes their accurate positioning, often at high speed, by sequences of binary bits, through whose means a traffic route (often a preferred route) is set up over available wire facilities from origin to destination. Parts of the switching circuits which set it up withdraw for like service on other calls. Other parts of the plant supervise the connection, often make prearranged interpolations or excisions of transmitted information, and at a proper time pull the circuit down, redistribute its portions and apparatus for subsequent use, and make records upon which to bill for services rendered.

As in the parallel case of dial telephony, the switching sequences are directly initiated by the sender (calling subscriber).[2] He may employ a push-button, dial, or teleprinter keyboard. Dials are employed on some European telegraph networks. Two-character director (selector) signals originated from teleprinter keyboards are the preferred fully automatic form in network telegraphy in the United States although ordinary push-button or supervisory key signals often suffice to bring in a central-office telegraph operator on a subscriber's call. Control of the starting and stopping of the motor in the customer's teleprinter by the central-office operator is often employed as a reverse signaling and control device in the initiation of telegram transmission.

By *director signals* supplied either by the patron or by the switching operator, the telegram is taken into tape reperforator storage or transmitted to destination. A second set of director signals may be employed to indicate priority handling, in which event they will activate a path to reperforator storage and deferment of subordinated messages, or to designate a special disposition of the message within the reperforator system.

The second vital element of control is the *knock-down* or *end-of-message indicator*, universally supplied from the teleprinter keyboard, which causes circuits to be restored to nonuse normal, either directly or through supervisory-cord signals. Together, the director and end-of-message signals mark transmissions off into message units, each one of which in sequence may be headed toward a different destination.

Taking tape into temporary storage affords the slight time delays which permit characters to be interpolated or extracted in response to other switching controls. For example, strings of messages from two circuits may be interleaved (flip flop) into one circuit under control of selector and knock-down circuits. Or a tape may pause, under control, to pick up a sequence number from a cam-driven automatic numbering machine. A director signal, having utilized its effectiveness in throwing a switching sequence into operation, may be reintroduced in an initial position for rereading at the next reperforator point (cross office or otherwise).[3] The time of day (A.M. or P.M.) may be introduced into the transmission from a coding device geared to a timepiece.[4] A translator may be interposed to change certain characters into different ones to render the sending keyboard compatible with the receiving teleprinter. Or tape may

[1] Keister, W., A. E. Ritchie, and S. H. Washburn, "Design of Switching Circuits," Van Nostrand, 1951. Caldwell, S. H., The Recognition and Identification of Symmetric Switching Functions, *Trans. AIEE*, **73**, Part 1, 142, 1954.

[2] Currie, F. L., Automatic Local Sending from Telephone Recording Positions, *WUTR*, **10**, 129, 1956.

[3] Fisher, W. H., Design Trends and Concepts as Applied to Private Wire Telegraph Engineering, *WUTR*, **10**, October, 1956.

[4] Edgar, W. S. W., Jr., The Automatic Time and Date Transmitter, *WUTR*, **2**, 150, 1948.

be edited under scrutiny and stopped for manual correction for one reason or another. Other examples will be found in connection with treatment of private-wire systems.

Since incoming and outgoing circuits usually operate at the same normal speeds, it is evident that, if any pause in transmission, however slight, is to take place within a reperforator office, the cross-office facilities there must operate at higher than normal speeds if a gradually increasing accumulation of unsent tape is to be avoided. This leads to *double reperforation* and completely loose coupling between incoming and outgoing circuits. The cross-office linkage is supplied by rotary switches[1] or crossbar switches,[2] actuated by reading and sequencing circuits or by manual push-button control.

In a reperforator office which is neither fully automatic nor semiautomatic (push-button-operated), the loose coupling between incoming and outgoing circuits is accomplished by tearing the incoming tapes between message units and manually feeding them, in proper sequence, into transmitters connected to outgoing circuits.

Timing circuits and slow-acting relays find considerable use for circuit supervisory purposes,[3] either to initiate alarms or to take direct action when, after a measured interval, some expected sequence fails to take place (usually through human delay or error) or some required choice has not been finalized.

Reading (director) circuits, involving relay trains wired to disclose the occurrence of given character combinations in the transmission, are used to initiate switching functions.

Sequence circuits[4] are arranged to take note of incoming events (like first-come first-served) or the switching requirements or limitations of outgoing facilities (such as a limited number of trunks), and to sequence the switching in accordance with formula.

Telegraphy has benefited from and contributed to recent progress in information handling by digital computers. The switching specialist treats his plug, jack, key, multipoint relay, multiterminal switch, diode, rectifier, varistor, and transistor problems of mechanizing intelligence by methods of switching logic, switching algebra, and a descriptive geometry. He uses simple but versatile two-valued circuits which sense, find, count,[5] read, accept, reject, hold, lock out, collect, memorize, repeat, compare, check, correct, translate, choose, compute, hunt, test, select, operate, control, transfer,[6] connect, separate, time, record, restore, disconnect, and forget.[7]

In addition to the primary separation of messages from each other by means of director and termination signals, secondary separation between parts of a single message can be achieved by reading circuits. For example, the "top-line" message identification only may be separated out and sent to bookkeeping or monitorial printers (less message texts), or addresses may be separated from texts for purposes of handling multiple-address messages (books).

34. Private-wire Telegraph Systems. Networks rented by business and government from the common carriers range in size from the simple teleprinter trunks described in Sec. 30 to industrial and government networks extending throughout the country.[8] The points served usually comprise (1) relatively few switching centers (usually one to three) interconnected by adequate facilities and (2) a large number of tributary points joined with one or another of the switching centers. Messages are

[1] Blanton, W. B., and F. L. Currie, Push-button Switching, *WUTR*, **2**, 119, 1948.

[2] Hutton, R. W., Type-7 Crossbar Selector, *Elec. Commun.*, **31**, 32, 1954.

[3] Demonet, E. A., and L. C. Bower, Jr., Electronic Timers in Telegraph Switching Systems, *WUTR*, **4**, 93, 1950.

[4] Bennett, W. S., Minimizing and Mapping Sequential Circuits, *Trans. AIEE*, **74**, Part 1, 443, 1955.

[5] Grosdoff, I. E., Electronic Counters, *RCA Rev.*, **7**, 438, 1946.

[6] Blanton, W. B., and G. G. Light, The Development of Western Union Switching Systems, *WUTR*, **4**, 38, 1950.

[7] Keister *et al.*, *op. cit.*

[8] Johnson, L., and R. C. Stiles, A Fully Automatic Teletypewriter Switching Center for Military Use, *Trans. AIEE*, **73**, Part 1, 27, 1954. Schultheiss, C. E., High-speed Teletypewriter Equipment for the Armed Services, *Trans. AIEE*, **73**, Part 1, 88, 1954. Le Vino, R. B., Research in Signal Corps Teletypewriter and Switching Equipments, *Elec. Eng.*, **74**, 798, 1955.

keyboard-manipulated only once, and then at point of origin; subsequent relays are through either repeaters or reperforators to teleprinters of final destination.[1]

Determination of the number and traffic-handling characteristics of the wire facilities to be employed in a projected private telegraph system and the number and location of switching centers is a solution of a problem of systems planning,[2] subject to adjustment after experience with an initial installation. The number and location of switching centers and their size can be decided upon by methods of cut and try, only after a complete 3-day sample traffic-load study and growth estimate has disclosed the daily and peak-hour loads passing between every pair of stations on the system,[3] and after these data have been related to the message capacity of the tributary lines. Small offices may be served on way-wire (multistation) circuits. Full circuit loads, justifying *direct* circuits between certain pairs and by-passing switching centers, will be revealed; these loads will be separately allocated and only the residue considered with all other fractional-channel loads. A first design attempt might assume two switching centers and their location. Availability of bunched circuit routes, their mileage, and cost would then be found from each outlying tributary station to one or the other of the switching points, taking into account the superior operation advantages of grouped circuits. A second design attempt might consist in raising the number of switching centers to three and reducing it to one. Although some national networks have around ten switching centers, the trend is to reduce them to one or two to obtain inherent advantages of minimum delay, maximum accuracy, circuit flexibility, and staffing economies which may override the costs of greater circuit mileage.

The number of trunks to be furnished to interconnect switching centers and whether they are to take a *straight-line* (backbone) or *ring* route are matters for prime decisions based upon known laws of efficiency for grouped operation[4] and upon the laws of probability applied to predicted *fall of load* and *holding time* of connection.[5]

At reperforator centers, switching may be done: (1) manually (*torn-tape transfer*), in which case the messages in received reperforated tape form are physically separated by operators and inserted in transmitters which will take them another or the final step toward destination;[6] (2) semiautomatically (push-button switching) by an operator who, upon observing the destination of each message, pushes a corresponding button to direct it across the office to the circuit which will take the message onward; or (3) fully automatically, handled, without intervention of operators, by tape-reading *selectors* (*directors*) which are responsive to selection (directing) characters (usually

[1] Dirkes, R. F., A Modern Reperforator Switching System (Plan 51) for Patron Telegraph Service, *WUTR*, **2**, 138, 1948. Bacon, W. M., The 81-C-1 Teletypewriter Switching System, *Bell Lab. Record*, **28**, 145, 1950. Harris, M. R., and R. L. Samson, The Development of Western Union Switching Systems, *WUTR*, **4**, 112, 1950.

[2] Bancroft, E. P., Telegraph System Planning, *Elec. Commun.*, **27**, 175, 1950. Coggeshall, I. S., System Aspects and Trends of Modern Communication (bibliog.), *IRE Convention Record*, Part 8, 51, 1954.

[3] Dirkes, *op. cit.*

[4] Bancroft, *loc. cit.*

[5] Molina, E. C., Application of the Theory of Probability to Telephone Trunking Problems, *Bell System Tech. J.*, **6**, 461, 1927. Kruithof, J., Rotary Traffic Machine, *Elec. Commun.*, **23**, 192, 1946. Wright, E. P. G., Behavior of Telephone Exchange Traffic Where Non-equivalent-choice Outlets Are Commoned, *Elec. Commun.*, **24**, 42, 1947. Rabe, F. W., Variations of Telephone Traffic, *Elec. Commun.*, **26**, 243, 1949. Riordan, J., Telephone Traffic Time Averages, *Bell System Tech. J.*, **30**, 1129, 1951. Strunz, G., and R. E. Hubley, Application of the Theory of Probability to Western Union Switching Systems, *WUTR*, **6**, 133, 1952. Hayward, W. S., Jr., The Reliability of Telephone Traffic Load Measurements by Switch Counts, *Bell System Tech. J.*, **31**, 357, 1952. Riordan, J., Delay Curves for Calls Served at Random, *Bell System Tech. J.*, **32**, 100, 1953. Frost, G. R., W. Keister, and A. E. Ritchie, A Throwdown Machine for Telephone Traffic Studies, *Bell System Tech. J.*, **32**, 292, 1953. Wilkinson, R. I., Working Curves for Delayed Exponential Calls Served in Random Order, *Bell System Tech. J.*, **32**, 360, 1953. Truitt, C. J., Trunk Requirements in Alternate Routing Networks, *Bell System Tech. J.*, **33**, 277, 1954.

[6] Sparks, S., and R. G. Kreer, Tape Relay System for Radio-telegraph Operation, *RCA Rev.*, **8**, 393 1947. Pessagno, T. S., A Simplified Telegraph Switching System Plan 111-A, *WUTR*, **9**, 55, 1955.

two in number) preceding each message to designate its route, or its single or group destination.

Choice of type of switching operation represents a balance between cost and adequacy to render desired service. Torn-tape transfer is least expensive and is used in some large installations but is best adapted to light circuit loadings and smaller networks. Important considerations are: (1) whether the customer has particular reasons for wishing to adopt priority classification[1] of traffic movement (because it is relatively expensive to switch on a priority basis except with torn tape); (2) whether the subscriber has great need for fanning out multiple copies (books) of the same text to several or many addresses, for torn tape is less well adapted to this task than the automatic systems.

Adoption of automatic and semiautomatic systems requires full duplex operation of all lines, to prevent tape accumulations and to simplify circuitry.

Of some importance is the form of message adopted and the associated routine steps for its handling. Operational routines to be followed should ensure that all messages start printing at the left margin on a new line, clear of the preceding message; that the attack (selector or director combination) is cleared for unambiguous switching; that no fault in transmission occur which would divide a message in parts or dissociate any verifications (confirmations, collations) or supervisory instructions from the messages to which they apply; and that the message is properly terminated with the specified (knock-down, disconnect) signal combinations. The specific form of message and operating routine will, of course, be different for the torn-tape, push-button, and automatic-selector methods.

In a small simple torn-tape system, the tape output of half a dozen incoming printer-reperforators is produced in a console which also contains a corresponding number of outgoing transmitter-distributors. An operator, standing before the console, tears off one message at a time from each incoming tape at a point between messages which has been clearly marked as part of the end-of-message routine. She then examines the message for its destination and places it in the proper outgoing transmitter-distributor. If, for any reason, she cannot dispose of the tape immediately (*e.g.*, because the line is half-duplexed and is operating in the opposite direction, or because another tape is occupying the transmitter) or for some reason must hold it (*e.g.*, because other classes of message have standing priority), she clips the tape in a tape holder bearing priority labels that facilitate her subsequent handling of accumulated tape in orderly manner.

A common characteristic of a push-button or automatic reperforator switching center is that each message is reperforated twice (in tandem). The first reperforation is to receive it from one of the incoming lines; to reveal its destination for push-button or selector-switching purposes; and to make it available for high-speed cross-office transmission to the second reperforator. The purpose of the latter is to place in an orderly arrangement all messages due to leave the office over a given outbound line. Cross-office transmission is at higher speeds than either incoming or outgoing lines to make up for switching delays; it is made possible by designing the cross-office (second) reperforators to receive all five unit pulses from across the office simultaneously instead of seriatim. Speeds of 75 to 150 wpm are customary.

Special circuits used in push-button and automatic reperforator switching systems include:

Multistation circuits (*way wires*)[2] with provision either for lockout of idle stations until a busy station has finished sending or for making automatic periodic contact with each station in turn and establishing, in rotation, outlets to the switching office for those stations offering traffic (pickup). In the reverse (delivery) direction, automatic starting of printers, in rotation, for only those stations to be recipients of traffic on hand.

Station-control circuits at a multistation line point, which, upon reception of a unique or group office designator, connect teleprinter equipment to the line, then cut themselves out of circuit so as not to be responsive to the same letter sequences in traffic.

[1] Hyde, G. A. M., A Long Distance Automatic Teleprinter Exchange with Manual Priority Services, *Trans. AIEE*, **58**, 375, 1939.

[2] Adams, M. D., The Development of Western Union Switching Systems, *WUTR*, **4**, 69, 1950.

Reading (*director*) *control* circuits, set to recognize combinations of given characters appearing in a given order and to take appropriate switching action implied by the combination. Action may be delayed by a *sequencing* circuit, which makes commencement of a second activity dependent upon completion of a first.

Director circuits, which determine the office outlet by reading incoming director signals, which test the availability of a reperforator-transmitter at the outlet, then route the message cross-office.

Link circuits (connector switches), multipoint rotary or crossbar switch circuits, which complete cross-office connections between incoming and outgoing lines.

Load-distributor circuits, composed of (1) transmitter finder rotary switches which seek out idle sending trunks among an interswitching-center trunk group, and (2) load-distributor rotary switches which function to connect in even fashion any one of the cross-office circuits, serving that trunk group, to a trunk found by the transmitter finder, thus by-passing busy channels.

Multiple-address circuits, which respond to special two-letter group director codes, then: (1) read seriatim the codes of destinations; (2) for each code in sequence, complete a cross-office connection, hold it, block the outlet against other use; (3) in the case of each such connection to another switching office insert a group director code indicating that multiple address handling is required; (4) send the message to all outlets simultaneously (*master-send* circuit).

Group-code circuits which find uniquely grouped multiple-address routes by means of a single director combination, dispensing with the lists of destinations.

Intercept (*spill-over*) circuits to reperforators, provided to receive messages (1) requiring special supervision such as priority, deferment, special attention, closed-office storage (*willful intercept*), or (2) involved in false calls and other switching troubles (fail-safe *automatic intercept*).

Flip-flop (*walking-beam*) sequencing circuits, which relate two instruments (such as transmitters) to a single sequential activity (such as sending to line) alternately. The knock-down (termination) of the unit of work being performed by the first activates the second, and vice versa. The transmitter loaded first is first to start.

Safeguards[1] for handling traffic without loss due to circuit interruptions, switching errors, or handling mixups have warranted the development of special expedients, among which are:

Avoidance of the letters T, O, M, V, in the first director signal, as being particularly susceptible of occurring during preceding stop interval, resulting in wrong switching; avoidance of H, the end-of-message indicator, and Z, the cut-in signal for a receiving-only teleprinter.

Alarms on intercept circuits; alarms to indicate line failure, switching-circuit failure, or delay, printer, and reperforator failure, failure of tape to feed, torn tape, low tape supply. In certain instances the alarm also causes the failed equipment to test *busy*, against further use.

Read-back circuit alarm, used following inconsistent comparison of received numbers with numbering machines.

No-switch alarm which sounds when a switching unit reads a director combination that has not been assigned for use.

Automatic supplying of end-of-message (knock-down) signal after 45 sec nonreceipt.

Diary teleprinters, which print, on a single record, top-line identifying data for every message switched through a reperforator center.

Among the specialized equipment or usages developed for switching systems are:

Reperforators, Sec. 44.

Printer-perforators, which translate into printed characters the holes punched in the tape, in either of two aspects: (1) on the marginal edge of tape, in register with the punched holes; (2) on chadless (hinged, partial perforations) tape, in the middle of the tape, six center holes behind the punched signals.

Perforator-transmitters (FRXD), a swivel combination the tape storage within which is reduced to zero after reperforation ceases, avoiding tape feed-out to clear the message.

Transmitter-distributors, combining in one instrument start-stop distributor and tape-transmission functions.

Tape editors, for notching the edge of tape at a point where it is desired to stop its transmission for the purpose of using a keyboard to make a correction or insertion in the text.[2]

[1] Blanton, W. B., and G. G. Light, The Development of Western Union Switching Systems, *WUTR*, **4**, 38, 1950.

[2] Wishart, R. S., and G. G. Light, Switching to Patrons' Teleprinter Tie-lines, *WUTR*, **3**, 51, 1949.

Automatic numbering machines, which enter a transmitting circuit at the beginning of each message and insert a consecutive three-figure line number and circuit-identifying characters.

Use of rotary switches to connect an incoming line to a first-idle outgoing reperforator, trunk, or outlet for local delivery copies.

***Line finders*—rotary switches used to concentrate the use of a few connector switches among many lightly loaded receiving circuits for the purpose of avoiding incoming reperforation.**

35. Teletypewriter Exchange Systems (see Chap. 24, Sec. 69). Teletypewriters are switched through central offices for temporary interconnection between calling and called subscribers on a time-charge basis in a manner similar to long-distance telephony. The service (TWX) is offered in the United States by the telephone companies[1] on a separate national manually switched network using telegraph circuits derived from telephone physical and carrier facilities. In Europe the several national systems of posts, telegraphs, and telephones offer similar interconnections (Telex), national and international, on a network generally using derived circuits, generally separate from telephone on the subscriber's premises, and generally manually switched. Some European countries have machine-switched Telex; in some the subscriber's tie lines have done and still do alternate duty for telephone and modulated-carrier telegraphy. The United States overseas radiotelegraph carriers offer interconnections between TWX in the United States and Telex abroad.

TWX installations in the United States feature supervisory circuits in switchboards[2] at the principal switching points, automatic subswitching centers controlled by switching points,[3] universality of keyboards, a directory of subscribers, uniformity of transmission, and a high degree of accuracy (see Sec. 20). Connections are half-duplex (to and fro) and a subscriber-to-subscriber connection is made before service begins to be timed. *Conference connections* may be set up involving more than two subscribers.

Telex in Europe[4] places emphasis upon use of a connection-confirming *"Who Are You?" key* built into the teleprinters, upon use of Baudot code, and upon operating routines. The teleprinters contain an auxiliary cam with axis transverse to the printer-selecting bars. Upon making a single rotation initiated from the distant end of the circuit, this cam will transmit to line the called subscriber's Telex identifying number. The calling subscriber, having been informed his connection has been completed, tests it out by signaling upper-case D, which has been reserved for the identity release signal. This signal will thereupon trip the cam release on the called printer, furnishing the calling subscriber with printed confirmation of correct connection and testing all parts of the two-way communication circuit and functions except the paper, ribbon, and type bars, and without waiting for the called party to come in.

36. Time-division Multiplex (TDM). Multiplex printer circuits, on which the utilized frequency spectrum is apportioned among channels by synchronously commutating brushes instead of by filter bands, have the feature of packing spectrum space more tightly than can carrier or facsimile, hence are used on limited-frequency circuits, like ocean cables, overseas radiotelegraphs,[5] and specialized land-line circuits, such as the *varioplex*, requiring a wider-than-teleprinter (23-cps) band (see also Chap. 24, Sec. 23).

[1] Duncan, J. A., R. D. Parker, and R. E. Pierce, Telegraphy in the Bell System, *Trans. AIEE*, **63**, 1032, 1944.

[2] Knowlton, A. D., G. A. Locke, and F. J. Singer, Switchboards and Signaling Facilities of the Teletypewriter Exchange System, *Trans. AIEE*, **55**, 1015, September, 1946.

[3] Watson, E. F., Fundamentals of Teletypewriters Used in the Bell System, *Bell System Tech. J.*, **17**, 620, 1938.

[4] Wilcockson, H. E., International Telex, *P. O. Elec. Eng. J.* (Part 1) **41**, 39, 1948; (Part 2) **41**, 98. Gibbs, D. T., From National to International Telex, *Telecommunications J.* (London), **6**, 17, 1954. Phillips, F. G., The New Telex Service, *Telecommunications J.* (London), **7**, 64, 1955.

[5] Callahan, J. L., R. E. Mathes, and A. Kahn, Time-division Multiplex in Radiotelegraphic Practice, *Proc. IRE*, **26**, 55, 1938. Hoover, R., The Application of Western Union Multiplex to Navy Radio, *WUTR*, **2**, 29, 1948. Peterman, W. C., and A. Mine, Four/two Channel Time-division-multiplex Telegraph System for Long-distance Radio Circuits, *Elec. Commun.*, **28**, 127, 1951.

A unidirectional three-channel five-unit multiplex faceplate has 15 segments in a ring, swept by a continuously rotating brush connected to line. For handling the five-unit Baudot code in unit sequence on each channel, the faceplate is divided into three equal sectors of five segments each, each sector serving one operator and taking away from his transmitter one letter of his message every time the brush makes one revolution. At the opposite end of the circuit there is a synchronous brush, a corresponding faceplate, 15 segments, a 120-deg sector, a printer, and a receiving operator for each channel.

It would be possible to arrange each faceplate with 21 segments, to each seven of which a start-stop teleprinter might be connected. But assuming that measures are at hand for maintaining synchronism of the brushes (their rotation never stops),[1] the start and stop pulses are superfluous and can be omitted, whence the first spectrum saving of multiplex over start-stop: the code is shorter, in ratio 5:7.

A second saving appears when TDMs use carriers for their line—that is, when channels on a pair of wires, made available by frequency division, are themselves time-divided (Chap. 24, Sec. 24). The three-channel 60-wpm 45-cps multiplex will fit as nicely into a 150-cps frequency slot as will one 60-wpm 23-cps start-stop teleprinter (at least for one carrier section, be it 100 miles in length or transcontinental). The reason is linked with distortion:[2] Multiplex's five units are each subject to total distortion (that is, late or early crossing of the reference axis in relation to perfect signals). So are teleprinter's seven units, including the stop pulse (which matters little if at all), but also including the start pulse, which is most important because if it is late or early its distortion is additive to the separate distortions of the five intelligence units, throwing them further off and destroying margin. Multiplex has no such trigger mechanism and is not subject to double jeopardy as described; and experience has shown[2] that output per telegraph carrier channel can be doubled and trebled with multiplex. Start-stop teleprinter has some overriding advantages which make it a favorite, but it requires more spectrum space than the multiplex, which in-phases on the train of received signals, with no lost time.

Where TDM methods are in use, the practical number of channels is either 1, 2, 3, 4, 5, 6, 8, or 12. Mechanical distributors have prevailed, but electronic versions have been proposed[3] and are in use.[4] Multiplexes may be paired for complementary opposite-directional full-duplex use.

When so duplexed they may be equipped for varioplex operation;[5] to a three-channel faceplate a machine switching unit may be used to admit a fourth pair (or more) of users, the channel pattern going to line being then changed from normal ABC-ABC-ABC-ABC to ABC-DAB-CDA-BCD. The 60-wpm speed originally available to each of three is thereby cut to three-fourths, or 45 wpm per channel upon entry of the fourth user, and would be further cut to 36 wpm upon entry of a fifth. Laws of chance and judicious assignment of as many as 60 pairs of users on subchannels prevent peaks of simultaneous use from occurring with any consequent impractically low transmission speeds.

The TDM principle is the basis of broad-band pulse-modulation[6] systems used in microwave applications for simultaneous telephony-telegraphy and in mobile tele-

[1] Turner, F. T., Communications Synchronizing Systems, *Trans. AIEE*, **72**, Part 1, 874, 1953.

[2] Cofer, T. F., Printing Telegraphy and Inductive Disturbances, *WUTR*, **1**, 48, 1947.

[3] Skellett, A. M., The Magnetically Focused Radial Beam Vacuum Tube, *Bell System Tech. J.*, **23**, 190, 1944.

[4] Hansen, T. A., and R. D. Slayton, An Electronic Time Division Multiplex Telegraph Set, *Trans. AIEE*, **70**, Part 1, 354, 1951. Bliss, W. H., and J. E. Ruedy, An Electron Tube for High-speed Teleprinting, *RCA Rev.*, **16**, 5, 1955.

[5] Holcomb, P., Jr., Basic Principles of the Varioplex, *Trans. AIEE*, **60**, 1102, 1941. Pierson, O. E., The Western Union Varioplex Telegraph System, *Elec. Commun.*, **22**, 101, 1944.

[6] Grieg, D. D., and A. M. Levine, Pulse-time-modulated Multiplex Radio Relay System, *Elec. Commun.*, **23**, 159, 1946. Landon, V. D., Theoretical Analysis of Various Systems of Multiplex Transmission, *RCA Rev.*, **9**, 287, 1948. Deloraine, E. M., Pulse Modulation, *Elec. Commun.*, **26**, 222, 1949. Hyneman, J. R., Pulse Modulation Systems for Telegraphy, *WUTR*, **3**, 69, 1949. Jacobi, T. E., Multichannel Radio-telegraph System for High-frequency Circuits, *RCA Rev.*, **9**, 704, 1948. Dow, O. E., A Time-division Multiplex Terminal, *RCA Rev.*, **13**, 275, 1952.

metering service. In that usage, extra signals are used for maintaining synchronism instead of phasing the receiver on the signal train. In wire applications, TDM is expected to develop usefulness in integrated data processing.

37. Nonlanguage Data Transmission. Fringes of common interest between business machines and communication lines are created whenever associated parts of office machines are required to operate in separated locations.[1] The teleprinter is an example: a typewriter keyboard separated from its typing platen. A multiplicity of typesetting machines may be controlled from a single distant keyboard.[2] Billing and listing machines can be dissected and their components separated for remote accounting.[3] Telegraphic methods may be applied to inventory control,[4] to ticket reservations,[5] to reordering of mercantile stocks; to the posting, display, filing, and producing of information on call (*access*, in the computer technique), and for the consolidation of statistical reports, compiling of payrolls, and issuance of vouchers, where distance between organizational units is a factor.

Teleprinter tapes and techniques are also used in the input, output, and memory systems of high-speed computers;[6] in the feeding of control information to lathes, milling machines,[7] textile machinery, and other forms of automation. In the factory there is a similarity between controlling individual machines, batteries of machines, and entire factory units in ganged operation, on the one hand, and group-director and master-send multiple-telegram methods, on the other.[8] Similarly, in the business office there is an alliance and demonstrated convertibility[9] between the punched holes of business-machine cards, as a result of which the storage, memory, entry, and access systems[10] of both may be combined in novel and useful forms of office automation;[11] *e.g.*, the sorting and tabulating features of cards may be combined with the flexibility of tape in interpolating stored or manually inserted information among items of new intelligence by telegraph selector-director methods (editing). Codes included in the tape will stop and restart the transmitter for keyboard insertions; deletions are made by stepping tape through transmitters having deactivated pins.

While most technical data-processing activity is being directed to the design of all these machines on a local basis (*i.e.*, intraoffice, intraplant), much thought is also being given to their integration: determining the role of intercity lines to feed to their inputs information and control signals and to distribute their outputs to points of utilization (integrated data processing—IDP).[12] The existing five-unit telegraph systems (public message, teletypewriter exchange, and private wire) are logical building blocks of integrated systems; six-unit, seven-unit, and eight-unit codes are obvious practicable extensions of five-unit characters in relation to the prevailing 150-cps and 170-cps carrier-channel separations; electronic time-division and counting techniques may be given an outlet of higher magnitude on half or whole v-f bands. Devices are readily contrived to send six-, seven-, or eight-unit characters over existing five-unit channels,

[1] Coggeshall, I. S., System Aspects and Trends of Modern Communication, *IRE Convention Record*, Part 8, 51, 1954.

[2] Smalley, M. N., Teletypesetter Equipment in the Bell System, *Bell Lab. Record*, **31**, 183, 1953.

[3] Ostline, J. E., Automatic Call Recording and Accounting in the SATT System, *Communication and Electronics*, no. 5, 56, March, 1953.

[4] Schmidt, E. L., and J. J. Connolly, Inventory System Using Digital Computer Techniques, *WUTR*, **5**, 148, 1951.

[5] Triem, W. R., and J. S. Jammer, Electric Ticket Reserving System, *Elec. Commun.*, **27**, 251, 1950.

[6] Wheeler, D. J., and J. E. Robertson, Diagnostic Programs for the Illiac, *Proc. IRE*, **41**, 1320, 1953.

[7] Cunningham, F. W., Controlling Machine Tools Automatically, *Mech. Eng.*, **76**, 487, 1954. Pease, W., An Automatic Machine Tool, *Sci. American*, **187**, 101, September, 1952.

[8] Casper, L., Textile Automation by Signal Control, *Trans. AIEE*, **74**, Part 1, 176, 1955.

[9] Dirkes, R. F., Telegraph Applications of Integrated Data Processing, *WUTR*, **10**, 96, 1956.

[10] Hollander, G. L., Bibliography on Data Storage and Recording, *Trans. AIEE*, **73**, Part 1, 49, 1954.

[11] Bibbero, R. J., Automation—A Survey, *Elec. Eng.*, **74**, 775, 1955.

[12] Meszar, J., Switching at Its Boldest, *Bell Lab. Record*, **32**, 441, 1954. Wright, E. P. G., and D. S. Ridler, Functional-diagram Approach to Electronics in Telegraphy, *Elec. Commun.*, **32**, 26, 1955. Horton, A. W., and H. E. Vaughan, Transmission of Digital Information over Telephone Circuits, *Bell System Tech. J.*, **34**, 511, 1955. Bashe, C. J., P. W. Jackson, H. A. Mussell, and W. D. Winger, The Design of the IBM Type 702 System, *Trans. AIEE*, **74**, Part 1, 695, 1955.

thus providing for the transmission of a useful admixture of administrative dispatches and nonlanguage data at will over present lines without substantial changes.

Paper tape and, recently, magnetic tape, prepared from source documents, like order forms, waybills, time cards, sales reports, or statistics, may be coded for access, transmitted over wires, and reused with stored information in the preparation of such other documents as management reports, production scheduling, inventories, periodic cash balances and payrolls, or to activate other data-processing equipment. Separation of data by classes is accomplished by selector codes. Tape so used is called *common-language tape*. In its storage and reuse, without manual intervention, but with deletions (by idling through) of obsolete and the addition of variable material, it is referred to as *self-perpetuating tape*. Multiple forms, tabulations, control of carbon and multilith copies, and operating procedures must be exactly prescribed to employ self-perpetuating common-language tape.[1]

38. Error-checking Systems (see Chap. 24, Sec. 8). Information handled without context makes extra demands upon the operating margins of communication systems, creating an ascending scale of *bit-accuracy* standards, from telephony at one extreme, through telegraphic plain language, machine encryption, pronounceable code, cipher, and figures, to data transmission at the other. Examination of the order of this listing will show descending order of redundancy in textual content, hence increasing difficulty on the part of the recipient to "read through" errors without securing corrections.

The *fixed-ratio* error-checking code used in radiotelegraphy is not employed on wires for reasons stated in Sec. 20. However, since accuracies of the order of 100 million characters per error are attributed[2] to the electromechanical systems of modern computers, a demand for their intercity integration by wire has given impetus to matching computer accuracy in telegraphy, inclusive of the terminal information-interchange apparatus as well as the lines. At a transmission rate of 100 wpm and an accuracy rate of 1 error per 100 million characters, it requires, on the average, 2 months of 10-hr days to develop one error.

Because of the complexity of their equipment, designers of computers have found it necessary, even when all operations are local, to incorporate *error-detecting* (and in some cases *error-correcting*) methods into their machines.[3] Reduced to terms of five-unit perforator-tape telegraphic transmission,[4] these involve the regularly spaced automatic insertion of extra "blind" characters into the tape which convey tally information regarding some statistical feature of the transmitted tape, against which the receiving reperforator can *check*—passing on those groups which do check but locking itself up on groups where transmission discrepancies are revealed. The group of intelligence characters between the check characters is designated a *word;* such words are of any convenient length—say, twenty five-unit characters, representing 100 binaries or an equivalent number of decimal digits. The number of characters in a *check group* might be four[5]—a burden of 20 per cent, equal to the adding of a sixth unit to the five-unit code. The nature of the check made by these four check characters is called a *parity check* of mark units transmitted in the 20 vertical columns and 5 horizontal lines of holes in the 100-bit *grid* under examination. "Mentally," the machine rearranges the grid into a 10 by 10 square; counts the marks first in each column, then in each line, assigning to each *even* count the parity value of M and to each *odd* count the parity value of S; finally punches these 20 M and S parity values seriatim in the 20 holes of the four-character test group.

[1] White, C. G., Railways Modernize with Centralized Supervisory Control, *Elec. Commun.*, **28,** 3, 1951. Brandt, W. A., and J. Polyzou, Voice-frequency Signaling Equipment, *Elec. Commun.*, **29,** 117, 1952. Dunn, R. H., and C. H. Chambers, Telemetering for System Operation, *Elec. Commun.*, **30,** 305, 1953. Smith, C. W., and M. E. Forrest, Jr., A Fully Selective Telemetering System Using Telegraph Facilities, *Trans. AIEE*, **74,** Part 1, 373, 1955.

[2] Gryb, R. M., Error Checking with Particular Reference to Telegraph Systems, AIEE Conference Paper No. 56-844, 1956.

[3] Richards, R. K., "Arithmetic Operations in Digital Computers," Van Nostrand, 1955.

[4] Hamming, R. W., Error Detecting and Error Correcting Codes, *Bell System Tech. J.*, **29,** 147, 1950.

[5] Gryb, *op. cit.*

At the final receiving reperforator the same tests are applied while the intelligence tape is being prepared there, but the 20 M and S parity values are stored, rather than punched, for comparison with the parity tally received over the line. If a transmission error occurs in the word under test, an *even* count (instead of *odd*, or vice versa) will show up as a discrepancy in an indicated column and on an indicated line, intersecting at the errored bit. Correction is obtained by retransmission of the errored character or errored word. The check characters having been automatically eliminated from the reperforated tape, the latter is ready to be fed, error-free, into the input of the computer. Such systems[1] are in the experimental stage.

Pertinent to the subject of error-checking systems are the subject and evaluation of *absolute line delay*, which become significant if *automatic error correction* (ARQ—Chap. 24, Sec. 71) is employed,[2] since it involves round-trip cumulative time to send by wire from A to B, receive and check any inaccuracy at B, seize the transmitter at B, switch the receiver at A, back-space at A, and repeat the character accurately from A to B. To lesser degree, absolute delay concerns *answer back* (Sec. 35) and other round-trip uses of equipment.

The figures of one-way absolute delay in milliseconds may be typically approximated for a 1,000-mile single section of carrier as shown in Table 5. These figures should not be confused with signal distortion expressed in milliseconds variation from ideal (Sec. 19).

Table 5. One-way Absolute Delay

Component of Delay	Delay, Millisec
Sending and receiving mechanisms	30
Carrier-vehicle band equipment (1 millisec per carrier section)	1
Line-propagation time, including repeaters (8 μsec per mile)	8
Total absolute delay:	
For a 1,000-mile 1-section carrier	39
For a 3,000-mile 2-section carrier	56
For a 3,000-mile 3-section radio beam	39

Source: Bramhall, F. B., Transmission of Business Machine Data over Standard Telegraph Channels, *WUTR*, **10**, 89, 1956.

39. Telegraphic Encryption. Privacy in telegraphic communication is in part safeguarded by the public's use of code and cipher and by the complexity of wire circuits and equipment employed. Special measures have to be adopted if it is desired to obtain a military level of security from wire-circuit surveillance or interception. These special measures are themselves military secrets. Some of those used in the past have been based upon published methods.[3]

A scheme used in the First World War and later may be illustrated as though, at the sending end, two ordinary perforated five-unit tapes were to be fed simultaneously into two distributor transmitters—one the intelligence tape, the other a key tape. As the successive characters were sensed by the two transmitters, a character from each tape at a time, and each of the corresponding five units of each character in sequence, transmission of unit polarities to line would be determined by identities in the two tapes (producing marks) or disparities between the two (producing spaces). Thus the intelligence units sent to line would not be those of either tape but would constitute a running report on the correspondence (or lack of it) of the two, hence unintelligible if intercepted without a key tape.

At the receiving end there would also be two local transmitters, one using an exact duplicate of the key tape, the other using a tape reperforated from the line's running report of unit correspondence. The correspondence signals would be used to modify the key-tape signals so that the original intelligence would emerge in the receiving cir-

[1] Slepian, D., A Class of Binary Signaling Alphabets, *Bell System Tech. J.*, **35**, 203, 1956.

[2] Sparks, S., RCA Electronic Multiplex and Automatic Correction Equipment, *Telegraph & Telephone Age*, July, 1951, p. 7.

[3] Vernam, G. S., Cipher Printing Telegraph Systems for Secret Wire and Radio Telegraphic Communication, *Trans. AIEE*, **45**, 295, 1926.

cuit and directly operate the teleprinter. The security of the system would depend upon the encryption adequacy, feasibility of duplication, safeguarding, and nonreuse of the key tapes.

40. Facsimile Telegraph Systems. Many thousands of photoelectric scanning, black-and-white dry-recording facsimile machines (see Chap. 23) are in use for local pickup and delivery of telegrams[1] over inexpensive v-f local wire pairs in cable. The useful city range without repeaters is about 5 miles in cable. A great deal of engineering has been done on the concentration of central-office and subcenter equipment[2] to permit a number of machines and pairs so terminated to serve several times their number of outlying machines. Many auxiliary devices and circuits have also been produced[3] to handle and discharge sending and receiving blanks to speed up the handling of such facsimile copy as telegrams and train tickets.[4]

There is no barrier in the technology against provision of telefacsimile exchange service with full automatic selector switching. For through circuits the transmission requirements are little more severe than for telephony;[5] for transmission involving storage en route, magnetic recording has been experimentally successful, though signal regeneration at repeaters has yet to be accomplished. In applying facsimile to intercity services, its v-f carrier-band-width requirement (1,336 bits per net scanned typescript character) handicaps it[6] in competition with teleprinters (7½ bits per character plus a frequency-division carrier guard-band loading factor of $^{150}\!/_{23}$, or about 50 bits per net printed character). Even vestigial-side-band transmission[7] narrows the discrepancy only in degree. Further cheapening of intercity lines, in conjunction with different yardsticks of tariffing such as per-page rates, might change the present outlook. With the exception of weather and military mapping, of tickets, and of pictures for the press, market surveys and experimental runs have not confirmed the existence of more imaginative uses of facsimile for transmission of graphic information than for typescript readily handled by teleprinters.

TELEGRAPH APPARATUS

41. Organizations of Apparatus. Most private-wire telegraph equipment is obtained from the catalogue offerings of the telegraph and telephone companies or specialty manufacturers, or assembled by making circuit adaptations involving such standard equipment. A principal item is the *teleprinter*, the model most widely employed being universally known as type 15, or, with perforator and distributor-transmitter accouterments (for *off-line* operation), as type 19. The most recent, and a higher-speed, printer is the type 28, a somewhat comprehensive description of which[8] is warranted here (Sec. 43). Another, but smaller, item is the *main-line relay*, whose universality entitles it to treatment from the viewpoint of design (Sec. 42).

Customers' equipment[9] is generally console-mounted, incoming and outgoing consoles being separately standardized. Printers are operated from a seated position; switching is done in a standing position. Consoles are unitized and made interchangeable, lending themselves to system expansion or contraction without change. Major equipment, packaged items, appliqués, subassemblies, and nodules are removable[10]

[1] Ridings, G. H., Facsimile Transceiver for Pickup and Delivery of Telegrams, *Elec. Commun.*, **26**, 129, 1949.

[2] Hill, A. S., Facsimile Systems (bibliog.), IRE Convention Record, Part 8, 24, 1954. Levine, S., A Simplified Telefax Concentrator, *WUTR*, **8**, 18, 1954.

[3] Cofer, T. F., Design of a Commercial Facsimile System, *WUTR*, **8**, 61, 1954.

[4] Zabriskie, D. M., Ticketfax Transmitters and Recorders, *Trans. AIEE*, **74**, Part 1, 536, 1955. Jelinek, C., Ticketfax Electronics, *Trans. AIEE*, **74**, Part 1, 565, 1955.

[5] Cannon, W. D., Delay Distortion Correction, *WUTR*, **10**, 58, 1956.

[6] Coggeshall, I. S., The Transmission of Intelligence in Typescript, *WUTR*, **7**, 61, 1953.

[7] Bramhall, F. B., Intercity Facsimile Trunks, *WUTR*, **8**, 57, 1954.

[8] Another American modern high-speed teleprinter is the Kleinschmidt, described by C. E. Schultheiss, High Speed Teletypewriter Equipment for the Armed Services, *Trans. AIEE*, **73**, Part 1, 88, 1954.

[9] Fisher, W. H., Design Trends and Concepts as Applied to Private Wire Telegraph Engineering, *WUTR*, **10**, 137, 1956.

[10] Cusack, F. H., Telegraph (Radio) Terminal AN/FGC-29 Equipment Features, *WUTR*, **9**, 10, 1955.

and, when in trouble, are handled on the basis of "swap-now-fix-later." Removal is facilitated by general use of multipoint base-jack or clip connections or multiconductor Jones plugs and jacks.[1]

Equipment units and wiring are made as accessible as possible by use of pivoted chassis or drawer construction. Flexibility is carried over into such items as a-c bus feeders and other power-supply wiring.[2] Extreme miniaturization and printed circuitry have not been widely employed in telegraphy. Much has been done, however, to attain compactness in units for mounting in standard 19-in. relay racks. One expedient adopted, for example, in design of filters for carriers, is to use more C and less space-consuming L in LC circuits.[3]

Provision is made for ventilation of apparatus and exhausting of room air to dissipate the heat losses encountered in apparatus. Cabinet blowers and room-air-conditioning units are fitted with air filters which give additional protection against accumulations of dust and lint on switching contacts. Removal of moisture concurrently with air cooling aids in the mechanical processing of perforated tape.[4]

Bay-mounted auxiliary equipment is set off to one side of a floor layout. Connections are provided to both wire and radio circuits by standardized transducers.

Repair benches and testing equipment are located adjacent to equipment installations. Some test circuits are wired in through cabling; in other cases it is advantageous to have test equipment mounted on dollies, to be rolled to position and there plugged in to test jacks. Portable, "rigidized" test equipment is furnished for common use of switching centers and offices tributary to them.

Military installations make special demands for shock mounting of rugged apparatus, imperviousness to moisture and mold (tropicalization) and to extremes of heat and cold, ease of transport (light weight, portability, and cabinet mountings), and tolerance of extreme power-supply voltage variations.[5]

42. Main-line Relays. Many thousands of relays are used in telegraph service, most of which are of leaf- or wire-spring type, generally known as Craft relays, even more prevalent in telephone-switching service. However, the telegraph main-line polar relay,[6] which follows plus-minus emf signal-arrival curves, is appropriately illustrative of principles of design.[7]

Polar Relays. These are more sensitive than single-current relays. In general two magnetic paths must be provided for the magnetic fluxes: one for the flux produced by the permanent magnet, the other for the flux produced by the operating current. These paths usually are arranged in the form of a bridge or similar balanced configuration so that a center position of the armature exists with respect to the permanent-magnet flux. The path for the flux produced by the operating current is disposed in the bridge in such a manner that the magnetic forces on the armature are unbalanced when operating flux exists so that the armature movement depends upon the direction of the operating current.

Figure 21 shows the magnetic conformations of a number of widely used polar relays. In the figure, each permanent magnet and its flux path are shown in heavy outline and the operating-current flux paths in lighter outline. The diagrams do not indicate relative dimensions or mechanical arrangement of the parts.

Actuating coils may be four in number: two main coils, which must operate within specified limits of current, and two auxiliary coils having half the number of turns of the main coils. The four coils may be poled and connected in various combinations. In one application, one auxiliary coil is caused to set up a force, as soon as the contact tongue is banked, say, on mark, to tend to help it leave for space upon decay of main-

[1] Klippel, W. H., A Test Set for Rotary Switch Shelves, *WUTR*, **3**, 101, 1949. Fisher, W. H., Apparatus for the Modern Reperforator Office, *WUTR*, **3**, 162, 1949.

[2] Fisher, *op. cit.*

[3] Taylor, R. C., Filters for a 150-kc Carrier System, *WUTR*, **2**, 166, 1948.

[4] Organ, N., Central Office Engineering of Modern Telegraph Offices, *WUTR*, **5**, 92, 1951.

[5] Cusack, *op. cit.*

[6] Courtesy of W. D. Cannon.

[7] Cannon, W. D., and T. Rystedt, An Improved Polar Telegraph Relay, *WUTR*, **6**, 29, 1952.

line current and somewhat in anticipation of full arrival of the next wave front; the other auxiliary coil is wired into a circuit which will energize it (hence, the armature) with a capacitor kick the instant the mark contact is actually broken. (The same coils work in reverse when the tongue is on space, giving double impetus to hasten its movement in response to a new mark arrival curve.)

In another application the auxiliary coils are wired into a vibrating circuit whose frequency matches that of bit reversals coming in over the line, whose energy is generated locally either by a tuned circuit or by auxiliary rings on the receiving distributor, and whose magnitude is equal to the received signal. Thus, the combined signals act to aid or to oppose, hence to regenerate the incoming signal (*cf.* Sec. 28, Fig. 14).

Relay Construction. In any relay, the flux reacts upon an armature which bears a tongue with contact points. The motion of the tongue between mark and space is not less than 0.006 in., but because of the foregoing auxiliary expedients its transit time (of the order of 1.2 millisec) is less than that governed by line current (10 to 35 ma, roughly 3 watts or +35 dbm), flux, and steepness of arrival curve alone.

Armatures should work freely throughout the life of bearings. Current-carrying bearings are minutely because of high current density through microscopic areas even when shunted by jumper wires, and are thus subject to seizing. Hole-and-sleeve pivots are therefore suspect; knife-edge or rolling-edge bearings are preferred; some modern relay armatures are spring-mounted and work without bearings.

Armatures may be mounted to provide two or three positions of the contacts. In the common two-position relay, the tongue rests upon either one contact or the other depending upon the direction of the operating current. In the three-position relay, the tongue stands midway between the two contacts when no operating current exists. This condition is usually provided by a spring mounting of appropriate stiffness to counteract the tendency of the permanent-magnet flux to hold the tongue on either one contact or the other.

FIG. 21. Magnetic circuits of commonly used relays: (A) Siemens (European); (B) Wheatstone; (C) type 17 Western Union; (D) Carpenter (British); (E) U.S. modification of (D); (F) type 215 Western Electric; (G) inefficient type—path of operating flux includes the permanent magnet (of high reluctance); (H) type 202 Western Union.

Usually some means is incorporated in relay structures to eliminate or minimize contact bounce; for example, compliance spring arrangements associated with the contacts are used to dissipate energy in sliding friction to arrest the motion of the armature. In type 202-A, a single spring is used for mounting and to implement the bounce-absorbing principle by dissipating the energy in sliding friction at the contacts. Usually a small bounce is unimportant since the commonly used *spark killers* (Sec. 17, Fig. 8*A*) across the contacts largely filter out such short-period variations.

To facilitate quick removal for replacement, adjustment, repair, or overhaul, main-line relays are universally base-jack- or plug-mounted.

Contacts. There is probably no universal contact material best suited to all specific applications and ensuring satisfactory operation and long contact life. The particu-

lar material selected depends upon many factors such as filming, pitting, and material transfer. Materials such as silver, palladium, and their commonly used alloys have excellent conductivity. Tungsten is resistant to mechanical wear but has considerably less contact conductivity. Tungsten carbide is free of many of the defects of other contact materials but has much higher contact resistance. Consequently tungsten carbide contacts must be operated at relatively high contact pressure. But because a relay equipped with tungsten carbide contacts will perform several billion operations before readjustment becomes necessary, they are favored for low maintenance cost and trouble-free operation in the 202-A relay and other types.

The importance of relays as circuit elements justifies references to the literature for further information.[1]

43. Model 28 Teletypewriter (see also Chap. 24, Sec. 69). Since the widely used 60-speed (368 operations per min) start-stop teletypewriter type 15 is described in Sec. 30,[2] the more recent, higher-speed type 28 merits description here.[3] Type 15 printers operate, with special attention, at 75-speed; type 28 is designed for 100- and will operate at 125-speed. Speeds of 60 to 66 wpm are well suited to the capacity of keyboard operators. Higher circuit (channel) speeds are made practicable if transmission is to be from prechecked, reperforated, stored, or otherwise accumulated tape. With accurate wire transmission, receiving puts no practical ceiling on speed, for printer copy requires only cursory examination by receiving operators, if not produced by unattended printers. The trend, therefore, is to higher circuit operating speeds.

There are nine code bars in the model 28 teleprinter (Fig. 22). The first, second, third, fourth, and fifth pulse code bars are positioned to marking or spacing, that is, to the left or to the right as viewed from the front of the teleprinter, by the selector mechanism. The positioning of the *common* code bar is determined by the position of the first and second pulse code bars, and has no other function than the vertical positioning of the type box.

[1] Curtis, A. M., Contact Phenomena in Telephone Switching Circuits, *Trans. AIEE*, **59**, 360, 1940. Fairweather, A., The Closure and Partial Separation of a Metallic Contact, *J. IEE* (London), **92**, Part 1, 301, 1945. Atalla, M. M., Arcing of Electrical Contacts in Telephone Switching Circuits (bibliog.), *Bell System Tech. J.*, **32**, 1231, 1953; **33**, 535, 1954; **34**, 203, 1955. Atalla, M. M., and R. E. Cox, Theory of Open-contact Performance of Twin Contacts, *Bell System Tech. J.*, **33**, 1373, 1954. Ellwood, W. B., The Measurement of the Transient Power and Energy Dissipated in Closing Switch Contacts, *Bell System Tech. J.*, **34**, 1191, 1955. Germer, L. H., Heat Dissipation at the Electrodes of a Short Electric Arc, *Bell System Tech. J.*, **30**, 933, 1951. Rudeforth, S., Contact Resistance and Its Variation with Current, *P.O. Elec. Eng. J.*, **42**, Part 2, 65, 1949. Gebert, E. B., and T. F. Cofer, Contact Materials in Telegraph Apparatus, *WUTR*, **5**, 32, 1951. Hermance, H. W., and T. F. Egan, "Fingerprinting" Relays (Contacts), *Bell Lab. Record*, **34**, 289, 1956. Gumley, R. H., Relay Contact Protection, *Bell Lab. Record*, **34**, 350, 1956. Hunt, L. B., and H. G. Taylor, Electrical Contact Springs, *J. IEE* (London), **92**, 38, 1945. Sumner, E. E., Relay Armature Rebound Analysis, *Bell System Tech. J.*, **31**, 172, 1952. Garbarino, H. L., and K. E. Bisshopp, A Polar Relay Using Momentum Transfer, *Trans. AIEE*, **74**, Part 1, 731, 1955. Shackleton, S. P., and H. W. Purcell, Relays in the Bell System, *Bell System Tech. J.*, **3**, 1, 1924. Keller, A. C., *et al.*, Design of Relays (259 pp.), *Bell System Tech. J.*, **33**, 1, 1954. Fry, J. R., Important Design Factors Influencing Reliability of Relays, *Bell System Tech. J.*, **31**, 976, 1952. Fry, J. R., and L. A. Gardner, Polarized Telegraph Relays, *Trans. AIEE*, **44**, 223, 1925. Jensen, K. L., The Design of a Polarized Telegraph Relay, *J. IEE* (London), **83**, 117, 1938. Turner, H. A., and B. Scott, A Polarized Relay of Improved Performance, *P.O. Elec. Eng. J.*, **43**, Part 2, 85, 1950. Cassidy, T. F., Manufacture of Western Union Type 202 Polar Relay, *WUTR*, **10**, 45, 1956. Keller, A. C., A New General Purpose Relay, *Bell System Tech. J.*, **31**, 1023, 1952. Nijs, L. J., and R. A. H. Faict, Special Purpose Relays, *Elec. Commun.*, **29**, 260, 1952. Markle, R. E., Development of a Wire Contact Relay, *Trans. AIEE*, **74**, Part 1, 742, 1955. Haglund, H. H., and A. W. Breyfogel, Thyratrons in Loaded Cable Multiplex, *WUTR*, **3**, 34, 1949. Simon, S., Some Applications of Cold-cathode Tubes to Switching Systems, *Elec. Commun.*, **29**, 207, 1952. Townsend, M. A., and W. A. Depp, Cold Cathode Tubes for Transmission of Audio Frequency Signals, *Bell System Tech. J.*, **32**, 1371, 1953. Brown, J. T. L., and C. E. Pollard, Mercury Contact Relays, *Trans. AIEE*, **66**, 1106, 1947. Brown, J. T. L., and C. E. Pollard, Balanced Polar Mercury Contact Relay, *Bell System Tech. J.*, **32**, 1393, 1953. Brown, J. T. L., Automatic (276-mercury-contact) Relay-adjusting Circuit, *Bell Lab. Record*, **33**, 60, 1955. Hovgaard, O. M., and G. E. Perreault, Development of Reed Switches and Relays, *Bell System Tech. J.*, **34**, 309, 1955. "Electronic Components Handbook," McGraw-Hill, 1956.

[2] "Electrical Engineers' Handbook," Electric Communication and Electronics, 4th ed., Sec. 18:5, p. 7, Wiley, 1950.

[3] Benjamin, A. S., A Step Forward in Printing Telegraphy, *Trans. AIEE*, **73**, Part 1, 10, 1954. Zenner, W. J., A New Teletypewriter, *Bell Lab. Record*, **31**, 262, 1953. Smith, F. W., Modern High Speed Teleprinters, *WUTR*, **9**, 110, 1955.

The *suppression*, *zero*, and *letters-figures shift* code bars are positioned to marking or spacing by functions in the *stunt box*.

The code bars have two separate and distinct functions. First, after having been positioned for a particular character by the selector mechanism, they cause the type-box positioning mechanism, by means of mechanical linkages, to position the *type box* with respect to the *printing hammer*, so that the character whose code is set up on the first, second, third, fourth, and fifth pulse code bars will be printed. Second, they cause typing-control functions to be performed within the printer mechanically, or at a remote location by means of switching.

Fig. 22. Code bar arrangement showing the five intelligence selector bars and auxiliary bars.

In the typing unit,[1] the type pallets are mounted in a rectangular type box approximately ½ in. thick, 1 in. wide, and 2 in. long. The weight of this type box and its associated *carriage* is only 8 oz, about one-tenth as much as the type-carriage assembly model 15. This permits the carriage to return from the right to the left margin within the time of two characters, even at 100 wpm.

The *type pallets* in the left half of the type box (Fig. 23) are letters and those in the right half are figures. When the type box is in position to print letters, a *printing-position pointer* in the center of the top of the letters half of the box is just below the position in which the next character will be printed. When a *figures-shift* signal is received by the typing unit, the type box shifts to the left so that the figures half of the type box is in the printing position.

Each half of the type box consists of vertical rows of type pallets, four on each side of the printing-position pointer, and four horizontal rows deep. The character to be printed is selected by moving the type box vertically and horizontally to move the type pallet into

Fig. 23. Type-box code chart showing location of type pallets.

the printing position. The four possible combinations of the first and second pulses are used to determine which of the four horizontal rows of type will be selected. The third pulse determines whether the type box will shift to the right or to the left from its rest position, and the fourth and fifth pulses determine how far it will shift from its rest position. Each of the four horizontal rows has a different combination of the first and second pulses. Each character to the left of the printing-position pointer has a marking third pulse and each character to the right of the pointer has a spacing third pulse. Also, the combinations

[1] Smith. *op. cit.*

of the fourth and fifth pulses are different for each of the four possible vertical rows on either side of the center position.

After the type box has been positioned, a spring-operated *print hammer* which moves with the type box is unlatched and allowed to drive the selected type pallet against the ribbon and paper to print the character. The print hammer is then reset to its latched position and the type box is returned to its rest position. When the signal for a nonprinting function is received, the type box is positioned in the usual manner and the print hammer is operated; however, *dummy* type pallets which do not print are placed in each of the nonprinting positions, so that no impression is made on the paper. A *spacing-suppression* mechanism prevents the type box and carriage from spacing when a nonprinting function is selected.

Another advantage gained by use of the type box, in addition to its light weight, is the ease with which the entire type box can be removed and replaced. For example, a standard *communication* type box can be removed and a *weather-symbol* type box installed in its place in a matter of seconds. Each character is mounted on a separate type pallet, so that it is not possible to print part of a letter and part of its associated upper-case character at the same time as in the case of type bars. Since the type box shifts to change from letters to figures, the platen remains stationary except during a *line feed*, thus reducing paper *creep* when multicopy paper is used and also eliminating the shock produced by shifting a heavy platen or type-bar assembly from the letters to the figures position. Type alignment is controlled in manufacture and it is never necessary to solder type pallets in position with a soldering fixture to obtain accurate alignment, as it is on a typebar teleprinter. When it is necessary to replace a defective part, the entire type box can be replaced by a spare without removing the teleprinter from service.

Selecting Mechanism. The principle of operation of the selecting mechanism is the same as in the model 15 with differences in detail of the cam-and-lever assembly. A *mark-space lock-lever* cam and five *selector* cams are released and start rotating when the start pulse, preceding the five units of intelligence, is received. An interlocking arrangement ensures that a selecting armature is unlatched only momentarily at the time of each signal transition so that it is free to respond to the incoming signals, whereupon the armature is locked in its selected position until just before the next signal transition is due to occur. Thus a virtually *terminally regenerated* signal is passed through the cam-lever transfer mechanism for the proper mark-or-space positioning of five selectively slotted *code bars*. In addition to the five intelligence transfer levers, there is a sixth or common transfer lever, which is contrived so that, if either the No. 1 or the No. 2 transfer lever moves to the marking position, the common transfer lever will also move to the marking position.

Fig. 24. Code-bar shift mechanism.

Shortly after the five selector levers have operated in sequence and set up the received code combination on the code-bar shift bars, a shift-lever link is driven upward by a code-bar positioning clutch and associated mechanism. As the shift-lever link rises (Fig. 24), the two code-bar shift levers engage notches on the code-bar shift bars and position them to the right or left depending on whether the shift bars are in the spacing or marking position.

To control the vertical positioning of the type box to print the selected character, the Nos. 1 and 2 code bars and the common code bar have projections on their right ends which extend beyond the code-bar guide (Fig. 22). In addition, there is a suppression code bar, marked "Sup" in Fig. 22, which extends beyond the code-bar guide. When the No. 1, 2, or the common code bar is in the spacing position, the projection on the bar will be in the path of point *a* on vertical positioning levers (Fig. 25). When one of these code bars is in the marking position, it will be moved to the left far enough so that the projection on the code bar will not be in the path of the vertical positioning levers. The projection on the suppression code bar, however, is always in the path of the vertical positioning levers.

Shortly after the code bars have been positioned, a type-box clutch and associated mechanism thrust a driving lever (Fig. 25) upward. This lever has attached to it an assembly consisting of the three vertical positioning levers. The type-box carriage track is fastened to the upper of these three levers, as shown in Fig. 25, so that the carriage track and type box move upward as the vertical positioning levers are driven upward. (There are actually two driving levers and two sets of vertical positioning levers, one on each side of the typing box, but only the right-hand set is shown.) As these levers move upward, point *a* on the middle lever strikes the projection on the lowest code bar which is in the spacing position. This causes the lever to rotate about its pivot *b*, and the lower end of the middle lever rotates counterclockwise so that the knee joint buckles at the pivot *c*. As the driving lever continues to rise, the two lower levers continue to buckle and the upper lever stops its upward travel. A latch (not shown) momentarily locks the upper lever in its uppermost position until after the selected character has been printed. Thus the combination of the first and second pulses received by the teleprinter determines which of the four horizontal rows of type will be brought into position to print.

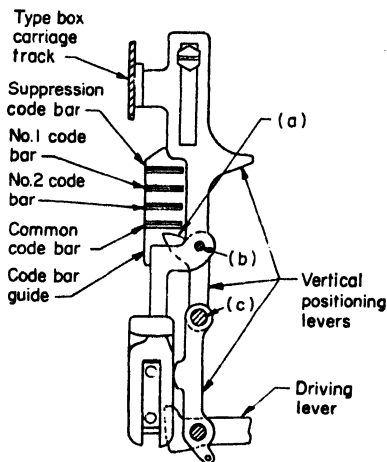



FIG. 25. Vertical positioning mechanism.

The position of the No. 3 code bar determines in which direction the type box will be shifted by determining which of two horizontal positioning drive links will be allowed to position the type box. When the No. 3 code bar is in the marking position, for example, a knee link which drives the type box to the left is buckled and another knee link which drives the type box to the right is allowed to operate and move the type box to the right. When the No. 3 code bar is in the spacing position, the knee link which drives the type box to the right is buckled and the other knee link is allowed to drive the type box to the left.

The combination of the fourth and fifth pulses received by the teleprinter determines how far the type box will shift horizontally from its rest position. The mechanism which accomplishes this operates on the same principle as the vertical positioning mechanism; that is, the Nos. 4 and 5 code bars both operate bell cranks which in turn operate horizontal stop slides. A common horizontal stop slide equipped with two stopping surfaces is operated by either of the other two slides so that four stop positions are obtained to correspond to the four possible combinations of the fourth and fifth pulses. These four stop positions determine at what point a horizontal positioning knee linkage will be buckled.

Two code bars not previously identified are shown in Fig. 22. The code bar marked *O* is part of the automatic carriage return and line-feed linkage and the one marked *S* is the letters-figures shift bar, the position of which determines which half of the type box will be in the printing position. These are controlled through a major assembly, the function box, which not only controls the typewriter "stunts" but has many collateral switching functions.

Function Box. This is located behind the code bars and extends across the entire width of the typing unit. Each of the code bars is notched at the rear in an identical pattern. These notches in the code bars are used in conjunction with function bars in the *stunt box* to select a function. When a particular combination for which a function bar is coded is set up on the code bars, the uniquely complete row of slots which represents it will first be sensed, then communicated by a bail-driven linkage to a function lever. The motion of this lever is used to initiate the desired function. The movement of the *lower end* of the lever is used to perform mechanical printer functions through appropriate mechanisms associated with the various function *levers*. A projection on the lower end of nonprinting function levers causes type spacing to be suppressed when a nonprinting function is selected. The upper end of the function levers is used to operate contacts for switching or to actuate other functions.

Stunt Box. There are 42 slots in the stunt box, each of which will accommodate a function bar and its associated mechanism. Some of these positions are required for such com-

mon functions as line feed, carriage return, word space, figures shift, letters shift, and bell signal; to these may be added horizontal tabulating, back spacing, automatic paper feed-out, sprocket feed, indexing to standard office forms, and similar mechanical elaborations of the typewriter as an office machine adapted to integrated data processing.

The remaining positions are available for other purposes. Since the function bars are coded (rather than the code bars), a function bar coded for any desired character, either upper or lower case, can be placed in any vacant slot in the function box. A function bar may also be coded to respond to both upper and lower case.

Some of the already explored[1] examples of versatility of the stunt box as a telegraph-operating adjunct are: (1) The function lever's switching contacts can be "make" contacts or "break" contacts. The combination of two sets of contacts, one make and one break, can be activated simultaneously to provide "transfer" contacts by means of activating two function-bar mechanisms on the same character. (2) A function may be performed only upon the reception of a stated sequence of two or more characters—that is, the function box may be used as a reading device to watch for and respond to given sequences of characters, like two-letter switching signals following a disconnect signal. (3) An electrical switch mechanism activated by a mechanical function in the stunt box can be used for remote control of some associated apparatus, like cutting in a reperforator upon receiving a particular character or a sequence of characters. In case of a sequence of characters, the last function-bar mechanism of the sequence will activate the switch mechanism. By adopting certain measures the reperforator may be left cut in so it can follow line signals. (4) By means of sequence-signal control of the suppression code bar the printer may alternately be thrown into "print case" or "stunt case." Among other things this makes it possible to select by single (if desired) call letters a number of teleprinters connected in series on a way wire. Stations *A*, *B*, *C* may be selected by those letters while printers are in stunt case; they will not falsely be selected by the occurrence of those letters in telegrams being transmitted while the printers are in print case. (5) For data-processing applications the stunt box can be equipped to control transmitters, reperforators, or other teleprinters. Data to be transmitted usually contain a great deal of constant material. This material can be stored in a master tape. When the master tape is placed in a transmitter, the stunt box can be used to stop and start the transmitter for manual insertion of variable data or insertion of data by another transmitter. The stunt box will also *edit* the information being transmitted by means of cutting in and out several reperforators with switching characters to select only the desired information that is to be retransmitted to various destinations. (6) By taking advantage of the machine's responsiveness to the suppression of printing or spacing, its versatility as a control mechanism is almost without limit.

44. Reperforators and Transmitters. The mechanisms common to a five-unit, Baudot code, telegraph-keyboard tape perforator and a five-unit teleprinter are similar mechanically interconnected grids, in configuration not unlike a crossbar switch, consisting of 32 actuating (in the perforator keyboard) or actuated (in the printer typing-mechanism) levers and, at right angles to them, five code bars which take up mark or space positions as a result (in the perforator) or as the controlling cause (in the printer) of the movement of the 32 levers.

Slots in and projections on the grid bars and levers are aligned distinctively, as the tumblers in a Yale lock, to admit of encoding (for perforating) or decoding (for printing) only one coded character combination at a time.

Whether the grid's bars and levers are to be moved by hand, by solenoids, or by motor-driven cams, are details,[2] as are their configurations. The fact that five-unit perforation can be controlled by 32 levers suggested that reperforation could be similarly controlled—that the code bars which selected the typebars of a teleprinter could also (even at the same time) select the punch pins of a perforator for the purpose of duplicating at station *B* the tape existing at station *A*. Such a duplicate tape could be switched for use in onward transmission, saving the labor of repunching it from a keyboard.

When combined with a teleprinter typebar-platen mechanism which, at the same

[1] Parkington, E. L., Function Mechanisms of the Model 28 Teleprinter, *WUTR*, **10**, 7, 1956.

[2] Harrison, H. H., Developments in Machine Telegraph Systems and Methods of Operation, *J. IEE* (London), **68**, 87, 1930.

time that it reproduces a tape, records on it the actual letters and figures represented by the punched holes (*i.e.*, a translation of the code), the reperforating device is called a *printer-perforator*. If the reperforator is combined with a tape transmitter for the onward electrical transmission of its output, the device is called a *reperforator-transmitter*. A combination of all three produces a *typing-reperforator-transmitter*.

Start-stop printer-perforators receiving directly from an incoming line[1] for push-button switching are more complex and expensive than cross-office reperforators[2] which work only locally and in which the translating function is unnecessary.

The third essential instrument, besides tape perforator and receiving teleprinter, is the (sending) transmitter, whose five sensing pins test the passing tape for holes (*marking*) or the no-hole (*spacing*) condition and throw five contacts either to mark or to space for transmission of the proper signals to line.

The speed of a transmitter is controlled by its outgoing or cross-office circuit, that of a reperforator by an incoming or cross-office circuit. If the outgoing speed is less than the incoming, an accumulation of tape takes place between reperforator and transmitter. If the outgoing speed exceeds the incoming, the tape will become so short that it will tear or break—unless a device, known as an *autocontrol lever*, which rides the tape loop between reperforator and transmitter, renders the transmitter tape-feed sprocket temporarily inoperative until more tape is produced.

When a message has been completely punched up in tape, perforation ceases, the tape loop tightens, and the autocontrol stops transmission, leaving the latter portion of the message in storage. If an operator is punching the message, he will insert a sufficiently long succession of nonoperative characters to serve as a tape feed-out, so that the last intelligence character of the message will pass over the transmitter sensing pins. If a reperforator is punching the message, it must sense the same situation and automatically insert tape feed-out characters. This introduces certain equipment complexities and extraneous characters which can be avoided by pivoting[3] the transmitter in such a way that as the tape loop tightens, the transmitter will "climb" it and will continue to do so until the last intelligence character has been sensed and transmitted. The first character of the subsequent message is immediately transmitted in such a typing-reperforator-transmitter-distributor. As manufactured, this device also contains a code-reading stunt-box appliqué unit not unlike that of the type 28 teleprinter; hence it has become one of the more versatile building blocks of private-wire switching systems.

45. Repeaters and Regenerators. A repeater is essentially a back-to-back juxtaposition of a receiver and a transmitter, inserted into a line for the purposes of dividing its length for speed advantage and of boosting attenuated power. Vacuum tubes or transistors in amplifiers are basic to the repeating of carrier and other modern forms of telegraphy including the narrow band widths of submarine cables.[4]

The significance of logarithmic attenuation with line length (Sec. 9) is recognized in the optimum spacing of repeaters in a line. Dividing the length of line l in half by insertion of a mid-way repeater increases its speed S and therefore its capacity to carry traffic four times (*i.e.*, by the square of the number of sections). This principle is expressed in an early submarine-cable formula, $S = k/CRl^2$. Line l and G, being small, are neglected; C and R have their usual values in per-mile units, and k is a speed constant which pertains to the apparatus and method of operation of the circuit.[5] It is not always convenient or possible to divide a circuit with repeaters placed at

[1] Blanton, W. B., and G. G. Light, The Development of Western Union Switching Systems, *WUTR*, **3**, 154, 1949.

[2] Blanchard, R. F., and W. B. Blanton, The Development of Western Union Switching Systems, *WUTR*, **2**, 2, 1948.

[3] Lake, R. A., Teletype Reperforator-Transmitter for Automatic Switching Systems, *Bell Lab. Record*, **26**, 106, 1948.

[4] Cramer, C. H., Submerged Repeaters for Long Submarine Telegraph Cables, *Trans. AIEE*, **70**, 566, 1951.

[5] Buckley, O. E., The Loaded Submarine Telegraph Cable, *Trans. AIEE*, **44**, 882, 1925. Coggeshall, I. S., Testing and Regulating Overseas, *WUTR*, **5**, 23, 1951.

optimum points. In that event the slowest section will determine the working speed of the whole circuit.

Another function of a repeater is to improve signal-to-noise ratio (P/N) by supplying power gain at a point in the circuit where the noise level is low. The telegraph-cable sea-bottom amplifier[1] is so situated, at a point shielded from extraneous noise, where a lift in the signal level permits it to override interference picked up subsequently in an exposed portion of the line.

Table 6 shows typical repeater spacing in miles for various kinds of line. The effects of noise ratios are evident in the different levels of gain suiting noise conditions encountered in shielded cable and on exposed open-wire line.

Table 6. Typical Repeater Spacings

Type of line	Max frequency	Transmission loss per pair-mile, db	Repeater spacing, miles	Amplifier gain, db
Submarine cable	20.8 cps	0.054*	2,148*	57*
Single open wire, grounded	25 cps	†	250	†
Single open wire, grounded	60 cps	†	250	†
Multipair cable pairs	30 kc	2.5	18	45
Multipair cable pairs	60 kc	3.0	15	45
Open-wire transposed pairs	30 kc	0.15	150	22.5
Open-wire transposed pairs	60 kc	0.2	100	20
Open-wire transposed pairs	150 kc	0.3	60	18

Repeater spacing on coaxial cables, from 5 to 7 miles at 1 Mc to 2½ to 3 miles at 10 Mc. All distances representative, but latitude is afforded in practice.

* Distances and losses in nautical miles. Repeater spacing dependent upon geography. Total loss of 115 db divided between one submerged and one shore amplifier.

† Repeater spacing empirical.

The simplest amplifiers are those which repeat a received carrier envelope with its carried modulations intact. At carrier terminals and other traffic "bundling points," where circuits converge or fan out, the carrier has to be demodulated into either v-f groups, individual v-f bands, separated subbands, or single telegraph channels; and while some of these will be terminated there, others will be repeated into other carriers and some into circuit legs and loops (Sec. 31), either with or without regeneration.

Regeneration may be accomplished by tube circuits[2] or by mechanical means. The latter involve back-to-back receiving-sending apparatus, consisting of synchronous brush-commutator (rotary) distributors and (sometimes on the same shaft) outgoing brushes which re-create fresh onward signals from distorted incoming ones (Sec. 36). A lag is introduced between received and sent signals, producing an interval during which pulses are held in storage (commonly as charges on capacitors) so that irregularities in the line *east* will not be repeated into the line *west*. Capacitor storage can be employed in the extreme case where signals from the line east may be absent, because of trouble, for several minutes, yet upon restoration the repeater will be found to be in step (synchronized).

On grounded telegraph land lines or on local loops, operated at 66 cps or less, repeating of signals is achieved purely by relay circuits. Repeaters may be one-way or duplexed. Single-line (back-and-forth) land-line repeaters require special continuity-preserving circuits. These were once of considerable interest and importance but now are all but obsolescent in an age of one-way carriers and four-wire lines. Modern d-c repeaters feature full flexibility in relation to their duplexed or half-duplex oper-

[1] Cramer, *op. cit.*

[2] Carter, R. O., L. K. Wheeler, and A. C. Frost, An Electronic Regenerative Repeater for 7½-unit Start-Stop Telegraph Signals, *P.O. Elec. Eng. J.*, **41**, Part 4, 222, 1949. Wilder, H. F., Electronic Regeneration of Teleprinter Signals, *WUTR*, **5**, 121, 1951. Pierson, O. E., A Transistorized Radioteletype Converter, *WUTR*, **9**, 64, 1955.

ation, and as to the number of loops and lines which may be brought into network arrangements through the use of dummy circuits.[1] The most widely used repeater (other than the interconnection of carrier sections) is the half-duplex network repeater. The reperforator-transmitter-distributor combination is also, of course, a teleprinter repeater of a loose-coupled type.

46. Functionally Switched Equipment. Under this heading may be grouped a number of kinds of shared telegraph equipment which have the common feature of being somewhat detached from individual telegraph lines themselves but which are switched in and out of circuit in the routine of handling telegrams. The list includes concentrators, reading circuits, directors, allocators, load distributors, timers, sequencing equipment, line finders; busy signaling, alarm, and test equipment; code translators; numbering and time-and-dating machines; and bookkeeping and monitor printers. Some of these functions are described elsewhere (Sec. 34). The purpose of this section is to point out this similarity: that a restricted number of pieces of auxiliary equipment are required, by switching methods often automatic, to cover a larger number of connected items of apparatus or circuits in turn. The engineering determination of the number of ancillary pieces so required can be calculated according to the principles of probability applied to holding time and group efficiency (Sec. 33, references).

A manually operated switchboard involves obvious relationships, of the type mentioned, between the numbers of lines, of calls per unit of time, of terminal instruments, of multipled panels, of patching-cord circuits, and of operators. A telegraph-line *concentrator*,[2] similarly, may serve 400 tape-teleprinter tie lines by means of 11 rotary switches, 44 central-office receiving printers, and 15 operators. In each group of 50 tie lines (balanced as evenly as possible), the first 5 calls go directly through a secondary line finder to the lowest-numbered teleprinter not in use; the next three calls, if the first 5 teleprinters are found busy, go through second, then through first secondary rotary switches to the lowest-numbered idle teleprinter. Since the lowest-numbered idle rotary switch always answers the next call, the calls received during light traffic periods will be concentrated on a minimum number of adjacent operating tables. It is also provided that the loads allocated by the switching system will be equalized among the operators until the point is reached, on an increasing load, that an additional operator should be assigned.

As rotary switches cut themselves in and out of circuit in response to received calls, so do code *translators* required to convert tape-teleprinter signals for page-printer operation.[3] For example, they are required to count the number of characters arriving in message texts and, upon receipt of the first word space following the fifty-eighth character, to (1) stop the transmission, (2) delete the word space, (3) substitute a carriage return and a line feed, (4) reset the counter to zero, and (5) restart transmission. For most conversions the translator adds characters, hence is a loose coupler between received tape signals and sent page signals. Other land-line wire conversions consist in changing the coded symbols for period, comma, apostrophe, per cent sign, etc., on account of keyboard differences between tape and page transmission.

One stunt initiated by a tie-line translator is to switch the entry of a *time-and-date unit* at the termination of a telegram.[4] After supplying automatically a notation such as "135 PM SEP 26 56" a rotary switch device transmits a carriage return and eight line feeds to provide spacing between successive messages. In the same way that this time-and-date unit obtains its encoded information from its memory (which is kept

[1] Morgenstern, R. V., Modern Telegraph Repeaters, *WUTR*, **7**, 71, 1953. Easterlin, P. R., A Switching System for Dispatcher Test Wires, *WUTR*, **7**, 13, 1953. Shanck, R. B., Flexibility of the 43A1 Carrier Telegraph System, *Bell Lab. Record*, **34**, 175, 1956.

[2] Morgenstern, R. V., and H. Buchwald, A Small Office Teleprinter Concentrator, *WUTR*, **8**, 72, 1954. Swanson, R. D., Plan 6 Receiving Teleprinter Concentrator in the Los Angeles Office, *WUTR*, **9**, 15, 1955. Howell, J. M., Receiving from Patrons' Teleprinter Tie-lines Plan 6-A Concentrators, *WUTR*, **4**, 30, 1950. Breyfogel, A. W., Western Union 50-line Desk-fax Concentrator, *WUTR*, **4**, 1, 1950.

[3] Frost, A. E., Tape-to-page Translator, *WUTR*, **3**, 61, 1949. Johnson, C. W., Concentrated Sending Position for Switching to Page-printer Tie-lines, *WUTR*, **9**, 143, 1955.

[4] Edgar, W. S. W., Jr., The Automatic Time and Date Transmitter, *WUTR*, **2**, 150, 1948.

up to date by clockwork), so a numbering machine[1] encodes consecutive number information, for use on the line, which it obtains from its system of Geneva wheels and cams.[2] Like many computer devices which "adjust" new to stored information, the numbering machine's entry into an informational train is governed by a reading circuit, awaiting a "trigger" combination; and its exit (and reversion to original transmission) is in response to its own last functional operation.

References

Specific references, appearing as footnotes, are not repeated here. Articles containing extensive reference lists are annotated (bibliog.). Also cf. references, Chaps. 24, 28.

Albert, A. L.: "Communication Engineering," Wiley, 1950.
Everitt, W. L., and G. E. Anner: "Communication Engineering," 3d ed., McGraw-Hill, 1956.
Fry, T. C.: "Probability and Its Engineering Uses," Van Nostrand, 1928.
Harrison, H. H.: "Printing Telegraph Systems and Mechanisms," Longmans, 1923.
Hausmann, E.: "Telegraph Engineering," Van Nostrand, 1922.
Herbert, T. E.: "Telegraphy," Pitman, 1930.
Hershberger, W. D.: "Principles of Communication Systems," Prentice-Hall, 1955.
Knowlton, A. E. (ed.): "Standard Handbook for Electrical Engineers," 8th ed., Telegraphy, McGraw-Hill, 1949.
"Reference Data for Radio Engineers," 4th ed., Federal Telephone and Radio Corp., 1956.
Smith, A. B., and Campbell, W. L.: "Automatic Telephony," McGraw-Hill, 1921.
Timbie, W. H.: "Basic Electricity for Communications," Wiley, 1943.
Ware, L. A., and H. R. Reed: "Communication Circuits," Wiley, 1949.
Wiener, N.: "Cybernetics: Control and Communication in the Animal and the Mach ne," Wiley, 1948.

General Articles:

Harris, L. H.: Telephony and Telegraphy—A Review of Progress (bibliog.), *Proc. IEE* (London), **101**, Part 1, 83, 1954.
Telegraph (Article), Encyclopaedia Britannica, 1950.

Monographs (Intramural, Available):

Babcock, W. C., E. Rentrop, and C. S. Thaeler: "Crosstalk on Open-Wire Lines," Monograph 2520, Bell Telephone Laboratories, New York, 1955.
Duplex Balancing—"Non-loaded Submarine Cables," Eng. Memo. No. 545, Western Union Telegraph Co., New York, 1949.
Keister, W., *et al.*: "Some Basic Relay Circuits," Bell System Monograph B—1688, New York, 1929.
"Principles of Electricity Applied to Telephone and Telegraph Work," American Telephone & Telegraph Co., New York, 1953.
Wilson, Leon T.: "Factors in the Transmission Characteristics of Cable Circuits," Coles Signal Laboratory, Western Union Committee on Technical Publication, New York, 1952.

Periodicals:

Bell Laboratories Record, monthly, Bell Telephone Laboratories, New York.
Bell System Technical Journal (*Bell System Tech. J.*), bimonthly, American Telephone and Telegraph Co., New York.
British Communications and Electronics, monthly, Heywood and Co., London, England.
Electrical Comunication (*Elec. Commun.*), quarterly, International Telephone and Telegraph Corp., New York.
Electronics, monthly, McGraw-Hill, New York.
Journal U.I.T. (Telecommunication Journal of the International Telecommunication Union), Geneva, Switzerland.
Post Office Telecommunications Journal, quarterly, General Post Office, London, England.
RCA Review (*RCA Rev.*), quarterly, RCA Laboratories, Princeton, N.J.
Western Union Technical Review (*WUTR*), quarterly, Western Union Telegraph Co., New York, N.Y.
Wire and Radio Communications, monthly, Carl A. Nelson, New York.

[1] Stebner, J. E., A New Automatic Numbering Machine, *WUTR*, **8**, 135, 1954.
[2] Blanchard, R. F., and W. B. Blanton, The Development of Western Union Switching Systems, *WUTR*, **2**, 75, 1948.

CHAPTER 28

TELEPHONY

By H. A. Affel[1]

INTRODUCTION

A telephone system intended to serve many individuals is necessarily complex and requires consideration of many factors. It must first permit setting up rapidly an exclusive channel from one party to any one of a large number of others which are remotely located. It must provide signaling or ringing to arouse the attention of the second party. When the connection is established, it must enable a two-way conversation to be carried on with privacy and with reasonable naturalness, volume, and

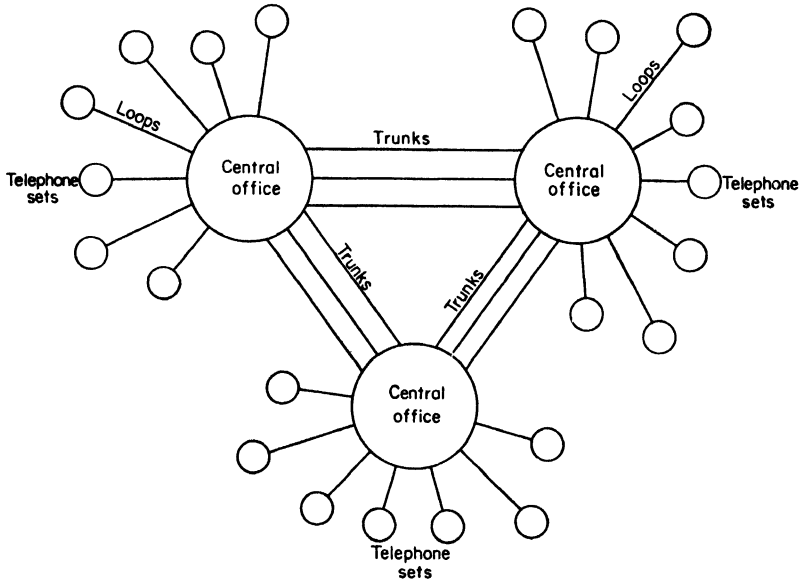



Fig. 1. Elementary telephone system.

freedom from unwanted noise. At the same time, economics plays a part, for all this must be accomplished at a price the users can afford.

1. Telephone-system Elements. The function which ensures that telephone users are connected properly to each other is generally called *switching*. The function of providing adequate communication paths of satisfactory speech quality is called *transmission*.

An elementary system is shown in Fig. 1. It includes (1) the *telephone set;* (2) a *loop* circuit connecting each subscriber to (3) a *central office* where switching may be accomplished; (4) *trunk* circuits for interconnections between central offices.

[1] Bell Telephone Laboratories, Inc.

In telephone systems extending over wide areas, several intermediate switching offices may be involved in order to afford flexibility and permit the interconnection of users nationally as well as internationally.

Some of the transmission paths may be made up of open-wire or fine-wire cable conductors. This is generally the case in loop circuits. Trunks may also consist of such facilities or they may include various types of radio links with or without carrier systems, in which case several communication paths are derived from one metallic or radio circuit. The choice of the communication link between the points involved will depend on the length, the numbers of circuits required, terrain conditions, and many other practical factors.

The switching systems, to permit the interconnection between a user and any other user, may consist of (1) a simple *manual* system with operators at switchboards at the central offices or (2) one of several types of *dial* systems in which an operator or the user himself sets into motion the equipment which automatically selects the desired telephone. In 1956 over 80 per cent of the telephones in the United States were dial.

In what follows, the different aspects of telephone-system design will be outlined and typical equipment considered in principle. References are given to sources of more detailed information.

2. The Telephone Set. A telephone set must have a *transmitter* and a *receiver*. A simple one-way communication system may, as shown in Fig. 2, consist of a carbon-microphone transmitter in series with a telephone receiver and a battery. The microphone has a diaphragm on which voice waves impinge and, by its movement, varies the resistance of a mass of carbon granules. The receiver consists of an iron diaphragm actuated by an electromagnet with a permanent-magnet core, which responds to the current variations created by the microphone. The diaphragm thus vibrates and passes on, more or less faithfully, a copy of the sound waves impinging on the transmitter.

Fig. 2. Elementary telephone set.

But a practical telephone set must embody other features, such as two-way operation and ringing. More components are required. Refinements are also needed in the design of the microphone and receiver. A typical modern set is pictured in Fig. 3. The various components are all mounted in the base of the instrument. Its circuit, shown in Fig. 4, includes, in addition to a transmitter and a receiver, various capacitors, a transformer or so-called *induction coil*, a *bell* or ringer, as well as switches and certain nonlinear elements whose function will be described later.

Fig. 3. Typical dial telephone set, 500 type.

In most cases the battery is located in the central office and supplies not only the power to the carbon transmitters of many subscribers but also plays an important part in supervision. This is referred to as a *common-battery* system.

In some older telephone systems and for special circuits generally not involving a central office, the battery may be located at the telephone set. A ringing generator, called a *magneto*, is then usually provided. Turning a handle of this permanent-magnet a-c generator provides a source of fairly high voltage which is applied to the line and actuates the bell at a distant telephone set or operates a signal at a switchboard.

3. Transmitter and Receiver. The modern transmitter and receiver offer higher stability and fidelity of transmission as well as efficiency as compared with models of the past. The simple disk of metal used as a diaphragm in earlier instruments tends to have a fairly discrete natural period of vibration, around 1,000 cps. It functions with maximum efficiency at this frequency and with reduced efficiency at other frequencies. Thus the speech transmitted is distorted, unnatural, and not of highest intelligibility. In both the transmitter and receiver of the new instruments, the diaphragms are no longer simple disks of metal but are shaped and supported to perform their functions more like pistons without sharp resonances. The airspaces in the set are also carefully proportioned to play a part in the over-all response.

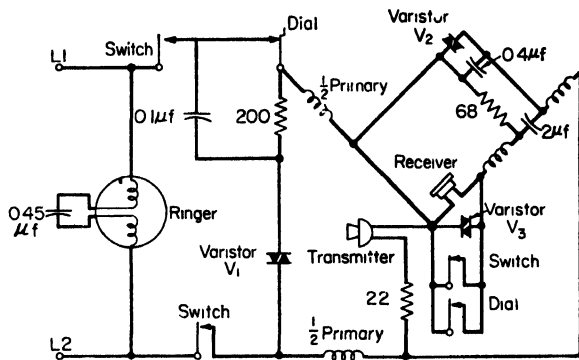
The carbon-granule transmitter has been greatly improved with time so that it is an instrument of comparatively high fidelity as well as stability and freedom from positional effects. Other types of microphones are, of course, feasible, such as an electromagnetic device like a telephone receiver in reverse, which was, indeed, Bell's original transmitter. Transmitters using piezocrystals of Rochelle salts or quartz, or condensers and other techniques, have also been employed for special purposes.

FIG. 4. Circuit schematic of 500D telephone set.

The carbon transmitter has, however, held its own in telephone practice because it is comparatively inexpensive and is actually a substantial energy amplifier as well as an acoustic-electric converter. Most other devices provide no amplification and are therefore comparatively inefficient. They generally require the interposition of amplifiers in the circuit.

The magnetic-acoustic design of the receiver has also been improved in recent years, and its efficiency as an energy converter is now fairly high, as well as reasonably uniform, over the required frequency spectrum. Cross sections of a typical modern transmitter and a typical receiver are shown in Figs. 5 and 6.

The transmission of conversational speech of reasonable naturalness and intelligibility requires the use of frequencies from about 200 to at least 3,000 cps. The gain in naturalness by transmitting lower frequencies and in intelligibility by transmitting higher frequencies is comparatively small and does not justify the considerably increased cost of the apparatus and system that would result from setting higher standards. The over-all frequency-transmission characteristics of a typical combined transmitter and receiver—sound path to sound path—are shown in Fig. 7 for loops of two lengths.

4. Speech Levels. In the case of the transmitter, when it is fed with a direct current of about 100 ma, the output of the telephone set is approximately −40 db below 1 volt across a 900-ohm pure resistance load with 1 *microbar*[1] input pressure at the lips of the talker. An average talker might develop a pressure of 20 db above 1 microbar, which would result in approximately −20 db below 1 volt or, roughly, 0.01 mw across a 900-ohm load. Loud talkers could readily exceed this value by a factor of 10, and the instantaneous peak voltages created might be another factor of 10 above the average.

[1] A microbar is 10^{-6} bar, which is equal to 10^{6} dynes per sq cm.

In the case of the receiver, 1 volt applied from a 900-ohm source will produce an acoustic pressure of about 30 db greater than a microbar at the ear of the listener.

5. Side Tone. The transformer or induction coil serves two main purposes: (1) by proper choice of winding ratios, it helps to establish optimum impedance relationships

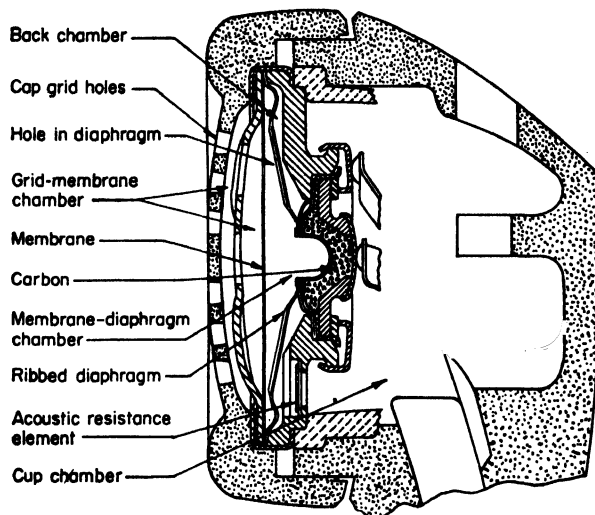



Fig. 5. Telephone transmitter.

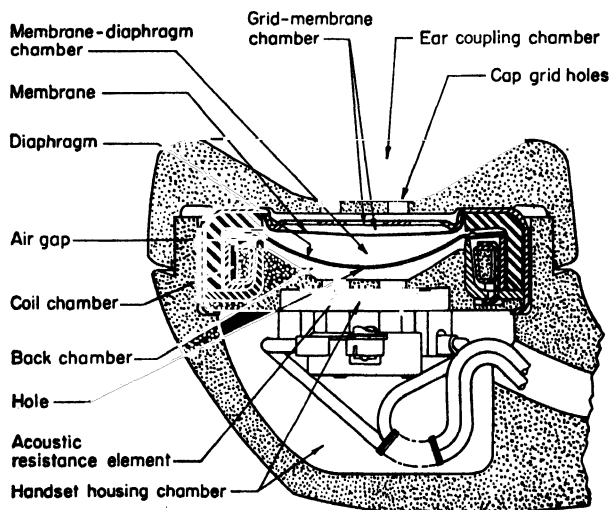



Fig. 6. Telephone receiver.

between the different elements of the telephone set and the line; and (2) it provides a degree of directional selectivity and tends to lessen the sound which returns through the receiver to the ear when speaking into the transmitter. Thus it minimizes the *side tone* when using the telephone set.

Excessive side tone is undesirable because it tends to discourage the talker from speaking sufficiently loudly into the transmitter. Room noises reaching the transmitter are also a possible source of annoyance. If the side tone is too great and there is an acoustic path between the receiver and the transmitter, either through the air or perhaps through the handle on which both are mounted, a feedback path may be established and sustained howling or singing set up.

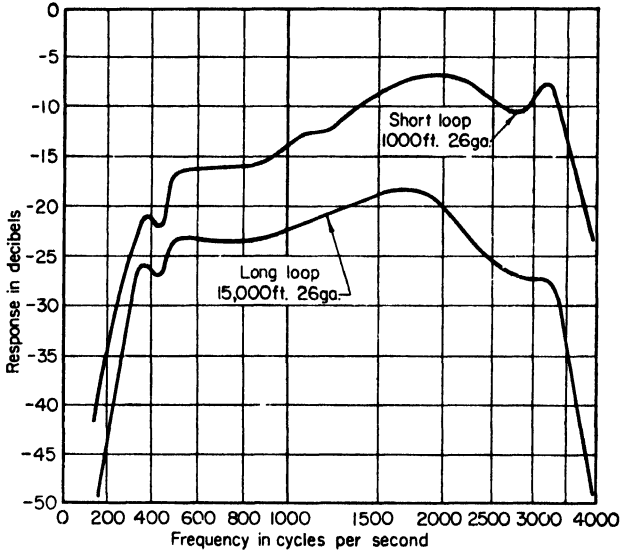



FIG. 7. Transmission characteristics of 500D type telephone set.

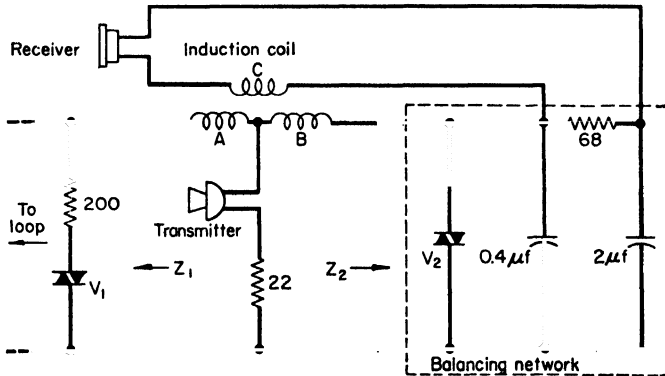



FIG. 8. Transmission circuit of the 500D telephone set showing side-tone balance.

The function of the induction coil in reducing the side tone can best be understood by reference to Fig. 8. It will be noted that the transmitter causes currents to flow through windings A and B, the former passing to the loop circuit and the latter into a *balancing network* whose impedance simulates that of the loop–central-office circuit. The A and B windings are arranged to have opposing effects on the core, and the C winding to the receiver therefore receives a minimum of voltage when the transmitter is actuated. Speech currents received from the central-office–loop circuit are, how-

ever, effectively transmitted directly from winding *A* into *C* with only a minor effect from the portion of the circuit through winding *B*.

6. Equalizer. Another feature of the type of modern telephone set shown in Fig. 8 is the *equalizer*. This consists of a resistor and nonlinear element whose function is to adjust the efficiency of the transmitter and receiver so that, if the telephone set is located at a considerable distance from the central office and has a higher-resistance loop circuit, its efficiency is greater than that of telephone sets nearer the central office. Thus the spread in effective transmission between different users is held to reasonable limits.

The nonlinear element is a silicon carbide varistor whose resistance decreases as current flowing through it increases. Thus, in a low-resistance loop, where the battery current is greater than in a high-resistance loop, it provides a greater shunting of the transmitter and receiver, and increases the loss in the circuit. There is also a varistor in the line-balancing network, whose effective resistance varies with the length of the loop. This tends to stabilize the balance and keeps the side tone from becoming too high for certain loop conditions.

7. Supervision. The telephone set commonly includes a *switchhook* which, among other things, closes the battery circuit when the handset is removed from the base. This powers the transmitter and actuates a relay and switchboard lamp at a manual central office to notify an operator that the telephone subscriber wishes to place a call. In the case of a dial system, the relay connects the user to a network of equipment which, like the operator, receives information about the called party that the user dials. The dial provides a means for sending accurately spaced (10 per second) pulses over the battery circuit to the central office. The dial functions are further discussed under Automatic Switching.

The calling signal is generally a bell with a polarized electromagnet structure whose hammer strikes a pair of gongs at each alternation of the ringing current. This is commonly 20 cps, as transmitted from most central offices. The set shown in Fig. 3 also has a means whereby the user can adjust the loudness of the bell.

The matter of ringing and supervision is complicated in the case of *party-line* service, where up to four and, in rural areas, sometimes more telephones are bridged on the same loop circuit. In some rural cases, ringing is still on a code-selection basis in which the bells at all parties ring for all calls. However, in most common battery-exchange areas, the ringing is selective and the bell rings for only the called party. This is usually done by making use of ground-return circuits provided by the two wires, also applying polarized pulses to obtain a total of four different combinations of ringing power. Gas-filled rectifiers are used in this connection in some telephone sets. In other selective-ringing arrangements, tuned-reed relays, responsive to particular ringing frequencies, have been employed.

8. Speakerphone. A comparatively recent innovation in telephone sets is the *speakerphone* or "hands-free" set in which a transmitter is mounted separately or in the base of the set and the receiver consists of a small loud-speaker which is located on the desk or table at some distance from the transmitter. Since both the transmitter and loud-speaker must include an amplifier to give them greater over-all efficiency, there is danger that the feedback path between the transmitter and receiver may provoke howling or singing—or, if not sustained singing, hollow-sounding speech.

Such a speakerphone set can, under most conditions, give hands-free communication if the proper precautions are taken in setting it up and adjusting the sensitivity of the transmitter and receiver. The acoustic characteristics of the surroundings also play a part. The most advanced designs of speakerphone systems have automatic voice-operated amplification control to minimize the feedback problem and ensure satisfactory operation under all circumstances.

9. Hard-of-hearing Set. Another auxiliary feature is included in a special set for hard-of-hearing individuals in which an amplifier is connected in the receiver circuit and mounted in the base, with the degree of amplification controlled by a knob operated as desired. The amplifier consists of a transistor in a simple circuit powered by

1 or 2 ma of current taken from the office common-battery supply which also actuates the transmitter and provides supervision. It seems probable that transistors will play an increasing part in the functioning of future telephone systems.

There are also available two modifications of the hard-of-hearing set. The first of these is intended for use in locations which at times are extremely noisy. This includes a push button on the set which, when actuated in an emergency, adds a substantial amount of amplification in the receiver circuit, at the same time cutting out the transmitter.

The other modification is intended for people whose speech power is greatly impaired. In this case, the amplifier, otherwise connected in the receiver circuit, is connected in the transmitter circuit to increase the volume which passes on the line. The amplifier is not used, of course, when the telephone is spoken into by a person with normal speech volume, as serious distortion would result from overloading.

10. Telephone-set Switching. Telephone sets can be provided with a limited number of keys or push buttons which permit the set to be connected alternately to different loop circuits, or for intercommunication with a limited number of other sets in the same office.

11. Recording and Answering. Recording and answering devices have been made available to function in the absence of a user. In general, these devices, installed on the user's premises, respond to an incoming call and play back a previously arranged recording, telling the calling party that the called party is absent and that a short-period message can be recorded if the calling party will dictate it. Or the responding mechanism may simply give to the calling party a more extensive message, not expecting him to dictate anything.

These devices consist of an association of relays, amplifiers, and either magnetic-tape recorders or, in some cases, mechanical-disk recording of the phonograph variety. A sketch of a typical recorder answering device is shown in Fig. 9.

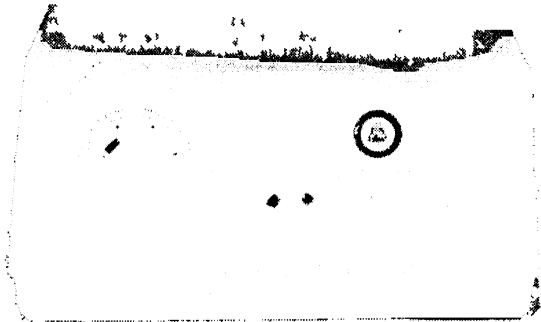



FIG. 9. Telephone answering set.

Another form of recording instrument is intended to be connected to the circuit and is capable of recording both sides of lengthy conversations between two telephone parties. If a record of this kind is to be made, the regulations of the Federal Communications Commission require that in some way both parties be made aware of the fact that the conversation is being recorded. This function is carried out by an attachment provided by the telephone companies, a *recorder connector*, which, every several seconds, applies to the line circuit a short spurt of 1,400-cycle tone of moderate strength. This has caused it to be referred to frequently as a "beeper" phone.

12. Sound-powered Telephone. For some special short-distance applications, particularly in the military field, use has been made of so-called sound-power telephone. These consist of electromagnetic devices which are like an efficient telephone receiver and can be used as either a transmitter or a receiver. Frequently two are mounted on the same handle, as in an ordinary handset. They have the advantage that, since the transmitter is a sound-electric converter, no batteries are required. In some cases, this provides a desirable element of reliability.

They have the disadvantage that, as a transmitter, the electromagnetic converter is

comparatively inefficient as compared with a carbon transmitter with battery. Therefore, they can be used over only short distances if no amplifiers are provided. The transmission distance using ordinary small conductors may be of the order of a mile or so compared with 25 miles or more for a carbon-button transmitter set. The exact comparison would, of course, depend on the size of conductors provided and the comparative line and room noise where the devices were used.

13. Protection. Close to where the user's loop circuit enters the premises, a *protector* is customarily installed to minimize the hazard to the user, the premises, and the telephone equipment in case the outside wires are struck by lightning or perhaps become crossed with a power line.

A protector basically consists of a gap between two carbon blocks having a separation of a few mils. Such gaps are connected between each line wire and ground. The most recent type of telephone protector, as shown in Fig. 10, consists of carbon block electrodes in a metal cap which, in turn, is screwed into a well containing a spring-loaded platform. On short-duration surges, arcing takes place only between the

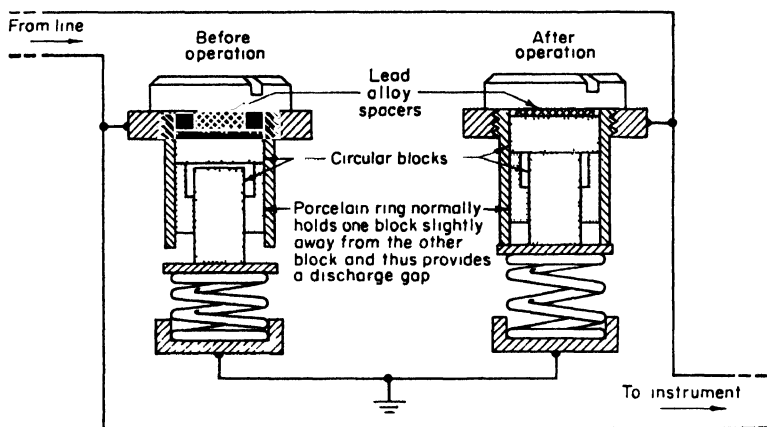



FIG. 10. Details of the 98A protector.

carbon blocks and the device returns to an open-circuited condition. However, on sustained arcing sufficient heat is developed to melt the alloy spacer which allows the spring to push the carbon blocks up into the well until the metal platform contacts the rim of the metal cap. Thus a low-resistance path to ground is established that is capable of carrying considerable sustained power current.

Some older-type protectors, which are still in use, also included the feature of providing a permanent low-resistance path upon experiencing a sustained flashover, but in this case, in general, two low-current high-voltage fuses were included in series with the two conductors between line and carbon blocks. The fuses would open the telephone circuit by melting the fuse wire and relieve the carbon blocks of the load of the sustained power currents. The newer protector of Fig. 10 has sufficient current-carrying capacity that, in general, the series fuses are not required.

In central offices, *sneak current* devices called *heat coils* are frequently used in addition to protectors. Heat coils are connected on the office side of the protectors and operate on small currents through the equipment to ground that may occur at voltages below the operating value of the protectors.

SWITCHING—GENERAL

It requires no stretch of imagination to picture a telephone system of many users in which each user is provided with an independent circuit to every other user, and a multipoint switch by which he may establish a desired connection at will. The com-

plexity and cost of such a system would obviously increase at a much greater rate than in direct proportion to the number of users and would soon exceed reasonable bounds. The efficiency of use, that is, the percentage of time the various interconnecting circuits were employed, would average very low in such a system.

The efforts of a telephone systems' designer are directed toward minimizing the number of interconnecting circuits, hence using them more efficiently. This calls for the establishment of central offices or other intervening switching points which are joined together by trunk circuits.

The telephone user originates the switching action. In manual operation, raising the telephone instrument places the user in contact with an operator at a switchboard in a central office to whom he gives the information by voice. In an automatic system the action of dialing passes on the called number to the equipment in the central office.

14. Switching Functions. Certain functions are required in a central office, either manual or automatic:

1. Means for identifying the calling line when the user raises his set from the switchhook
2. Means for making a connection to the calling line
3. Means for picking out an available transmission path to the called party, who may be connected to the same switchboard or to some other distant switchboard reached over one or more trunk circuits
4. Means for determining whether the called line is already in use
5. Means for signaling at the called station
6. In many cases, means for recording the location and duration of the call
7. When the parties at both ends terminate the conversation and hang up their sets, means for restoring all circuits to a state of readiness for other calls

In the case of a connection from one part of the country to another passing through several switching points, the design of the system must, as noted in Sec. 37, ensure that the over-all transmission loss, when the several links are connected together, does not exceed a reasonable value.

Also, the systems engineer must be certain that sufficient trunks are provided between switching points so that, in so far as economically practicable, the demand during the busy hours will not tie up all the available facilities and cause excessive delay in establishing a connection. This requires traffic studies by specialists who, with statistical information and probability theory plus experience, are able to predict how many trunks are needed between particular locations.

Included in such system design is, of course, the provision for alternate routing of connections in case of failure of a link in the preferred routing of a call or where the facilities are temporarily overloaded. This can be accomplished to a certain extent by operators in a manually operated system but, with an automatic-switching system, this function is carried out in a matter of seconds.

15. Private-branch Exchange. In the over-all switching plan, a loop circuit generally connects a user's telephone set to a central office where the first step in switching takes place. However, in many situations, as in organizations having large numbers of telephones, the first switching point may actually be on the premises of the organization in the form of a *private-branch exchange,* commonly termed *P.B.X.*

All the individuals' telephones are connected to this P.B.X. From it to a central office a certain number of trunk circuits are provided, substantially less in number than the total number of telephones connected to the P.B.X. Such a P.B.X., which may be either manually operated or a dial system, also permits intra-P.B.X. calling between the various individuals connected to the P.B.X. Thus the trunk circuits to the central office are used only on outgoing or incoming calls. P.B.X. installations of several thousand telephones are not uncommon in many large organizations, although they range upward from a comparatively few telephones in the case of a small organization.

16. Line Concentrator. Another primary step in switching may be provided by recently introduced equipment termed a *line concentrator.* Where there are many

unrelated users in an area located at some distance from a central office, instead of providing individual line circuits to the central office, such a concentrator may be installed near the group of users. It is an automatic-switching device into which all the users are wired, and which provides a limited number of trunks, say about one-fifth of the number of users in the group, between the concentrator and the central office. Thus it saves mileage in the loop circuits involved.

The relay equipment in the line concentrator, in conjunction with coordinating control equipment at the central office, automatically picks out an idle trunk when a user wishes to place a call. A concentrator can be considered as a sort of elementary central office or perhaps better as an extension of part of the central-office functions to a remote location. It is evidently also related in function to the automatic P.B.X., although it does not ordinarily provide means for the connected users to talk with each other without employing the trunks to the central office.

17. Components of Switching. Just as in transmission, there are various *devices* to carry out desired functions, such as inductors, resistors, and capacitors, so the switching designer also has various components at his disposal. These are mainly variants of the simple switch, *i.e.*, a means for making and breaking a circuit.

Some of these are, of course, quite familiar, such as the plug and jack. Where switching of one circuit is required to any one of a number of others, an operator may handle a flexible cable, termed a *cord*, on the end of which is mounted a *plug*. This may be thrust into a *jack* and thus make connections involving from one to three wire conductors. These are referred to as the *tip*, *ring*, and *sleeve* conductors. They derive their names from their position on the plug. The jack has springs which contact the tip, ring, and sleeve sections of the plug when the connection has been established.

Where a number of jacks are mounted together on the same structure, this is a *jack strip*. Many jack strips in close proximity provide a *jack field*. In turn, together with other equipment, they may make up a *manual switchboard*.

The other equipment on a switchboard consists mainly of *keys*. These are switches capable of being manipulated by an operator to make connection between certain desired portions of the circuit, including the cord. A key may have many springs or contacts. The key commonly used has an intermediate position as well as left and right positions to carry out its functions.

FIG. 11. Typical multicontact wire-spring relay.

A switchboard may also have small *signal lamps*, usually mounted close to jack positions, to indicate when a telephone user, whose circuit connects to a particular jack, wishes to alert the operator.

18. Relays. Perhaps the most important basic device of switching, both manual and automatic, is the *relay*. There are literally thousands of different types of relays, some, of course, differing from one another only in the number of contacts or *springs* which are affected.

The type of relay currently being applied in large numbers is shown in Fig. 11. This is a so-called *wire-spring relay*. It derives its name from the fact that the contacts are at the ends of springs of highly resilient wires. These wire springs take the place of flat springs which made up the structure of some earlier types of relays. The wire-spring relays have advantages in speed, sensitivity, and economy of manufacture.

19. Switching Systems. In automatic switching there are various special forms of switch-relay construction, for example, the *step-by-step* switch which appears in many different specific forms. In general, a step-by-step switch functions, as the name implies, to step a moving contact from one point to another each time a pulse is

received from the line circuit. Sometimes such step-by-step switches are arranged to select the proper contact point on a two-dimensional basis. The stepping mechanism first proceeds to find the necessary location in a vertical column and then moves at right angles to find one of a number of contacts across the row. Figure 12 shows, schematically, a particular form of step-by-step switch.

Another special switch connector is the *crossbar switch.* In this switch, shown in Fig. 13, a connection is made at a crossing of so-called *bars* in a rectangular-coordinate system. The bars in one coordinate are operated by a series of relays; correspondingly,

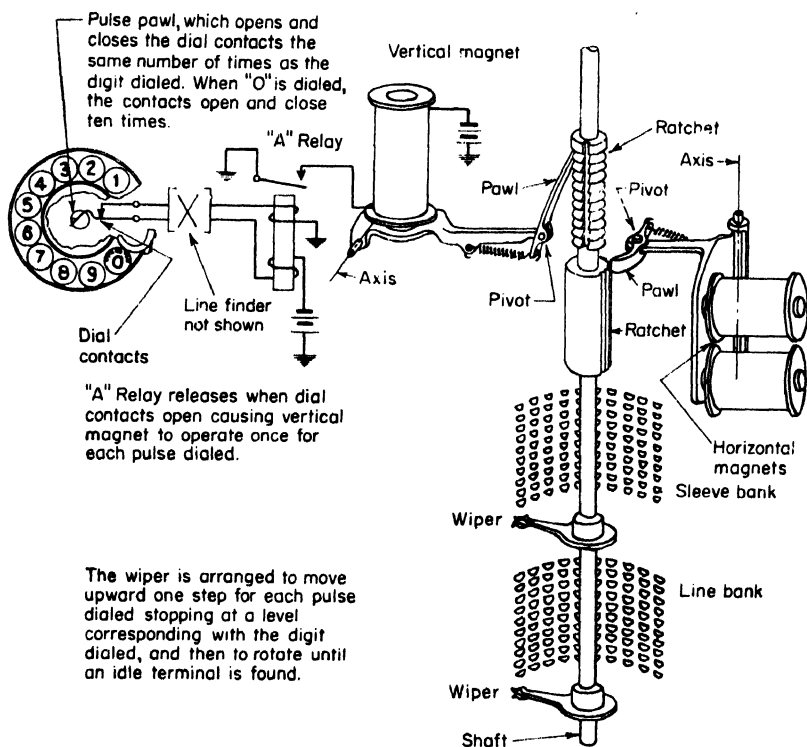



FIG. 12. Operation of step-by-step selector.

the bars in the other coordinate are responsive to other relays. Such a crossbar switch may make a connection at any one of 200 cross points, where there are, in effect, 10 relay bars in one and 20 in the other coordinate. A number of different cross-point connections may be made simultaneously. Each connection may involve two or more contacts.

Another common adjunct of all telephone central offices is the distributing or *cross-connecting frame,* into which all the circuits entering the office appear on *terminals* or lugs. Nearby, on other terminals, also appear all the circuits leading to switchboard or dial-switching equipment in the office. This permits running cross-connecting wires between the sets of terminals to tie together the office equipment and the line circuits in order to carry out certain layout arrangements. They may be changed from time to time by simply removing the old and attaching new cross-connecting wires. The whole arrangement is a means for providing flexibility and enabling the telephone office to meet new or changing service requirements.

20. Electronic Components. Even a brief summary of switching components would not be complete without mention of various electronic devices which promise to play an important part in switching systems of the future. One or two electronic devices have received application in the past, such as the cold-cathode gas tube which permitted eliminating the polarized relay in four-party ringing circuits.

Gas-filled or vacuum tubes can obviously be used in many types of circuits demanding fast relay-type action. The newer semiconductor devices, such as transistors and diodes, show great promise because of their small size, low power requirements, and speed of operation as compared with electromechanical devices, like relays.

For the storage of digital switching information, certain cathode-ray-type structures can be employed. *Magnetic-tape* or drum devices have been used experimentally. *Magnetic-core* devices, employing alloys of various types, including *ferrites*, have also been tested, as well as devices employing *ferroelectric* materials, such as barium titanate. It seems likely that switching systems of the future will be revolutionized in details of construction even though the broader principles of operation may continue to follow current practices.

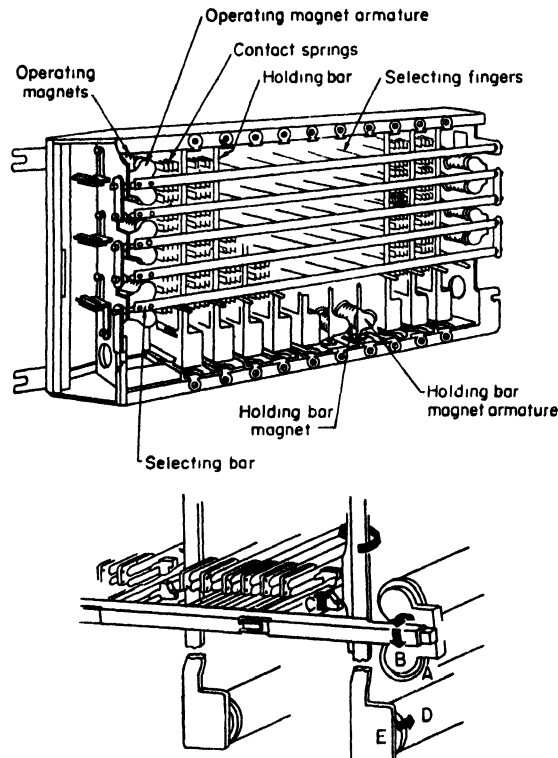



FIG. 13. Partial perspective of crossbar switch with 10 vertical units. Electromagnet *A* pulls projection *B* which tilts the horizontal bar so that a wire *C* projecting from the bar into the switch mechanism is raised into position. Magnet *D* now operates and pivots the vertical bar *E* inward. Bar *E* pressing against wire *C* closes the contacts and completes the connection.

21. Telephone-office Power. To operate the various equipments in a telephone office requires substantial sources of power, usually direct current. It is essential that they have great reliability. Storage batteries of substantial capacity are provided, and these are kept on floating charge through rectifiers or motor-generator sets driven from commercial power sources. Direct current at 48 volts is required in the greatest quantity, although other voltages up to 130 d.c. are also needed for special purposes, such as the plate battery supply for vacuum tubes. The storage batteries provided are usually of substantial capacity to take care of contingencies and in ample time to permit maintenance attendants to reach the office to start an emergency gasoline or diesel-engine set. Each telephone office is equipped with one or more such motor-generator sets which can carry the full office load indefinitely in case of a commercial power failure.

22. Manual Switching. The switchboard in a typical central office makes accessible to a single operator several thousand individual jacks for the various subscribers'

lines, as well as other jacks which connect to other central offices and to distant switchboards.

In an office of some size, these jacks appear in multiple at several operators' positions. Each jack has associated with it a signal lamp which lights on an incoming call. Thus several lamps in the multiple positions light simultaneously, and the first available operator seeing the light inserts the plug of her *answering cord* into the particular jack, whereupon the line lamp automatically goes out at all multiple positions. Figure 14 shows a cross section of a manual-switching-cord circuit. After the operator has

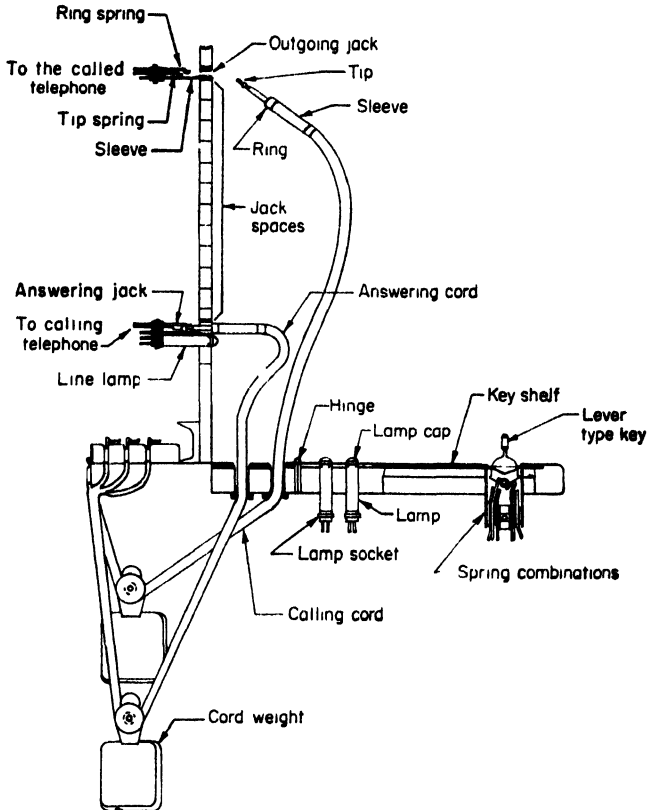



FIG. 14. Cross section of typical manual switchboard.

received the desired number from the user, she plugs the *calling cord* into the jack of the called telephone. An operation of the key switch applies 20-cycle ringing current to sound the bell of the called party.

In the case of larger central offices, the functions of receiving the calling signal, making contact with the user, and completing the call to another telephone in the same office are usually divided between two switchboards—the so-called *A* switchboard and *B* switchboard. The *A* operator, who has accessible perhaps 1,000 or 2,000 jacks, answers the call and passes on the called number to the *B* operator, who is able to reach directly as many as 10,000 or more jacks on the *B* switchboard.

If the called party is in another central-office area, the operator plugs the calling end of the cord into an interoffice trunk jack which signals another operator at a *B* switchboard who, in turn, makes the connection through her calling cord to the called party.

In the case of a call to a distant point, the initial operator plugs the calling cord into the jack connecting to another switchboard particularly adapted for handling such calls, and the telephone user gives the special operator further information about the call. This operator makes a record of the information and completes the call through other jacks connected to remote points. The call may be routed through other intermediate offices, since obviously all such switchboards do not have immediate access to other similar switchboards in the country. The originating operator also stamps on a ticket the exact time the called party answers and also when it is completed as indicated by the extinction of a light when either party has disconnected.

23. *Toll Dialing.* In the United States a large percentage of the intercity connections are now made automatically by an operator who dials the distant point, and the intermediate offices are automatically selected by the mechanism. In a growing number of areas in the United States, a user may dial another telephone directly in certain other areas without the interposition of any manual operator. The trend of development and application is definitely in the direction toward a time when all telephone users will be able to dial directly most telephones in any section of the country. Additional equipment is required for this purpose and it will be a few years before the dialing of long distance as well as local connections is universal. Of course, person-to-person calls will always require the assistance of an operator.

In anticipation of this eventuality, the United States is divided into areas each of which has a three-digit designation, which will be dialed first in connection with a long-distance call. Also, all users' telephone numbers are to be placed uniformly on a seven-digit basis, *i.e.*, two digits for the first letters of the central office, a third digit for the central-office number, and four digits for the individual telephone number. Thus, for example, a call from some distance to a Chicago number might be dialed 312 (area designation), then BUckingham 1-6200.

AUTOMATIC SWITCHING

In automatic-switching systems, functions otherwise performed by an operator are all automatically controlled by the dial impulses sent out from a telephone set. In current practice there are two important types of automatic systems: (1) the so-called *step-by-step* system and (2) the *crossbar-switching* system, both of which systems derive their names from the basic types of switches which are used to perform the operations.

The step-by-step system is becoming more and more limited to application in cases involving a small or moderate number of subscribers. It is a system of comparative simplicity, but large amounts of equipment are used exclusively for the duration of each call.

In the case of the crossbar system, which is a so-called common-control-type system,[1] much equipment of considerable complexity is involved momentarily in setting up a call, but it is not held for the duration of a particular call and so is available for setting up other calls. This general philosophy of design makes for efficient and speedy operation in large central offices.

24. Step-by-step Switching. The functioning of a step-by-step switch has been shown in Sec. 19. When a user picks up a telephone, the fingers of a particular step-by-step selector switch, functioning as a *line finder,* sweep through different lines and pause on the particular line where the telephone set has been lifted. Thereupon a dial tone informs the user that it is proper to make the call.

The circuit from the user is then connected to the fingers of the *first selector.* This selector, which is shown in Fig. 15, steps up the contact level corresponding to the first dialed letter or digit. The 10 terminals on this level lead to other selector switches. The selector automatically rotates across the horizontal terminals until it finds an idle terminal which leads to a second selector. As the second letter or digit is dialed, the *second-selector* switch steps up to the corresponding level. The same general process,

[1] An earlier so-called *panel* switching system, still used to some extent, also has common-control features.

in turn, leads to a third selector which completes the connection to a trunk to the desired central office in the area.

A similar step-by-step selection process involving three switches then takes place at the central office of the called party. At this point, with the numbers whose first two digits correspond to the numbers of the particular party dialed, the second selector rotates until it finds an idle terminal when a connection is made to the *connector switch.* The connector switch steps up to the level corresponding to the third dialed particular number and pauses until the fourth digit is dialed, whereupon it rotates the contact corresponding to the fourth digit, which is presumably the terminal of the user's line desired.

In the case of both the connector and selector switches, there is actually a second finger and set of contacts which operate mechanically in parallel with the talking-path contacts and which are used in connection with determining whether the called line is

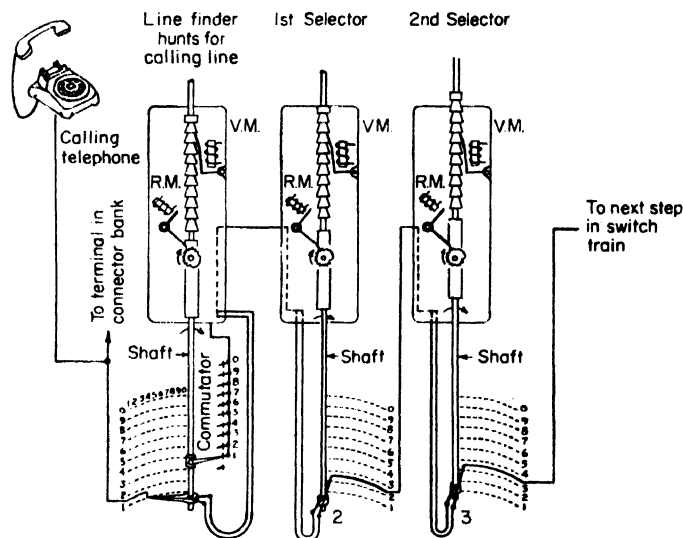



FIG. 15. Path of a call in step-by-step system.

busy or idle. When the finger has reached the terminal of the desired line, an automatic relay arrangement applies ringing current in timed spurts until the telephone is answered or the calling subscriber hangs up.

This description has thus far traced a call to a party in another central office. Actually, in the general case, with seven-digit subscribers' numbers which are now standard, a call to a party in the same central office is carried out by the same selection process.

In simple P.B.X. step-by-step switchboards, it is possible to get along with fewer steps where only a relatively few lines are involved.

25. Crossbar Switching. Crossbar switching differs chiefly from step-by-step in that the equipment does not respond to step-by-step signals as the successive digits are dialed. In a certain sense, it functions more like an operator who receives the call and then routes it in accordance with a knowledge of the circuits and circumstances involved.

There are two main divisions of equipment in a crossbar-switching system, which have different functions: (1) the control equipment which establishes talking paths by causing the proper switches to operate; (2) the switching network, mostly crossbar switches, through which the talking paths are set up.

To begin with, a group of crossbar switches can be used to make up a **crossbar-switching network.** As shown in Fig. 16, each small square represents a switch. A group of eight switches represents those which might be mounted in a frame in a central office. To simplify the drawing, only two lines come from each side of a switch, although in an actual switching central office, 10 lines would extend from each. The switches are evidently interconnected in a definite geometric pattern. If, for example, it were desired to connect points A and B, one could choose either of two routes through the network. In a central-office system, where 20 vertical-unit crossbar switches, such as previously pictured, are interconnected in a similar pattern, there are 10 possible electrical paths between A and B.

Fig. 16. Basic switching-frame arrangements in crossbar system.

A typical crossbar central office of 10,000 lines contains 1,800 crossbar switches and some 70,000 relays of all types.

26. Control Equipment. Figure 17 shows diagrammatically in more detail what happens in a No. 5 crossbar system. When the user picks up a handset, he is actually connected to a crossbar switch in a *line-link frame* of the switching network. In a fraction of a second a so-called dial-tone *marker* circuit is signaled. Quickly thereafter it sets up a connection from the calling line to an *originating register*, whereupon the marker is disconnected to handle other calls. At this point dial tone is transmitted to the telephone caller, who then dials the number desired.

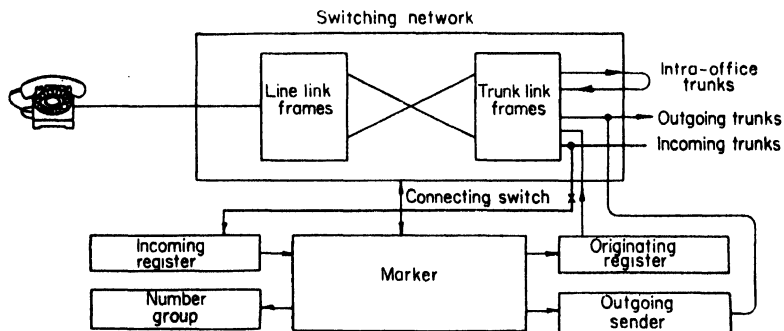



Fig. 17. Path of call in crossbar system.

The marker is a complex arrangement of relays which acts as the dispatching brain of the equipment. It "marks" points in the switching network (like A and B in Fig. 16) to which calls must go and selects and establishes paths between these points. This complex operation is performed by the marker within ½ sec per connection.

The originating register is a sort of telephone operator who passes back dial tone instead of saying "Number, please" and remembers the dial numbers, using relays and switches instead of a human brain. *Incoming registers* perform the same function for calls from other central offices.

When the last dialing impulse has been received, the originating register connects to an idle completing marker and transmits the number desired to it. The first three digits of the number, the so-called office code, tell the marker the name of the office called. If the called number is in another central office, the marker disconnects the originating line from the originating register, connects the proper *sender* to the proper outgoing trunk, and connects the originating user's line to the trunk through the *line-link* and *trunk-link frames*. The sender then transmits the number desired to another central office.

In the second central office, an incoming register records the number from the distant sender, after which it connects to a marker. The marker receives the number from the register. The marker now locates the called number in its office. To do this it checks the *number-group frame*. This is a mechanical telephone directory which knows the number of every telephone in the central office—knows to which terminals, on which line-link frame, each telephone is connected and is prepared to give out this information any time it is asked by the marker.

The marker obtains the location of the number, the right line-link frame, the crossbar switch, etc. Then the number-group frame and incoming register are dismissed by the marker, which now connects the incoming trunk to the called number through a trunk-link frame and line-link frame. Since the connection is now complete, the marker disconnects itself and goes about other business. The incoming trunk rings the called line and the call is complete when the called party lifts the receiver. At this time the users are connected only through the line-link frames and trunk-link frames in their respective central offices.

Senders are required in handling calls to other offices. Senders transmit numbers in special codes other offices are set to receive. Different types of senders are provided to transmit signals to different types of central offices, such as step-by-step, crossbar, and manual offices. In step-by-step connections, the codes are generally in a decimal system. If a different code is required, the senders and other switching equipment are prepared to translate from decimal to this code.

As simplified previously, the line-link frame is a unit consisting principally of crossbar switches to which telephones are connected. The line-link frame is the first link in the call from the user and the last also, since all telephones are connected to a line-link frame. A trunk-link frame is a unit to which trunks in the central office are connected. A trunk which begins and ends in the same central office is called an *intra-office trunk*. Lines or circuits which lead to or from other central offices are generally called *interoffice* trunks.

It is evident that in detail the equipment of a dial system is extremely complex—beyond the scope of this handbook to describe in detail. For example, a marker has about 2,000 relays. Fortunately, the marker as well as other office equipment which is relatively complex is not needed during the whole period of the conversation. They carry out their function in a fraction of a second, and the whole call is put through in considerably less time than it takes to describe even the major steps in it.

The functions carried out by the various major elements in the crossbar-switching system may eventually be carried out by electronic devices, such as transistors, diodes, electromagnetic memory units, and cathode-ray memory tubes. The chief advantages of doing this electronically are still greater speed of operation, smaller space for the equipment, and perhaps somewhat lower cost.

27. Intercity Switching. In a general sense, the functioning of intercity switching carries out more specifically and in greater detail the type of switching operation which starts with the originating office. However, as noted previously, three more digits are needed to direct the call to the particular region where the call is destined. Crossbar switches, relays, and, to some extent, step-by-step switches make up the switching equipment.

An important additional unit in the intercity-switching terminal is the *card translator*. This has some functions not unlike the number-group frame. It is a mechanical directory of routes which the call must take in passing from one area to another.

The information in this device is carried on large metal cards, as the name implies, which are punched in various ways. The punched holes permit light to pass through or be obscured, and these light beams actuate phototransistor units which, in turn, give the switches the information desired.

Intercity calls may be dialed by the operator directly if the called telephone is in a dial area or, as noted previously, in the case of certain selected areas, station-to-station calls may be dialed by the originating user.

28. Message Accounting. In the case of certain manual systems, where the accounting basis requires it, a *message register*, which is a counter associated with each user's line, is actuated each time the operator concerned with a call presses a register key. This enables the telephone company to keep track of the number of calls. In a dial system there is a similar arrangement which, in its elementary form, not only registers calls in the immediate user's area but also, if necessary, registers two or even more call units each time a call is made to extended areas which are a greater distance from the user.

More complex equipment is being installed, particularly in larger cities throughout the country, which is termed *automatic message accounting*, or A.M.A. In this equipment, information is registered by punchings on a tape for each call. This information includes the calling number, the called number, and the date and time of the beginning and end of the call. This accumulated tape for an office may be collected at periodic intervals and processed through automatic computing machinery which finally makes out a statement which can be billed to the telephone user. The tape processing and billing can be carried on at some central accounting center for a large city where information from several central offices is collected.

In some cities the automatic-message-accounting equipment associated with several central offices is located at the same point and the tape records are thus conveniently available for further processing. This is termed centralized A.M.A., or C.A.M.A.

TRANSMISSION—GENERAL

To enable persons to carry on a satisfactory telephone conversation depends on a good many factors. Some of these are under the control of the telephone-system designer; others are a function of the user's habits, such as:

How loudly the talker speaks, and how he holds the transmitter with respect to his lips

The noise conditions at the listener's telephone location

How good the listener's hearing is and how he reacts to different volumes; also how he holds the receiver with respect to his ear

29. System Transmission Parameters. Of those items which are under the control of the designer, the telephone set, as already discussed, is very important. This includes the conversion efficiency of the transmitter and receiver, their frequency transmission characteristics, side tone, and other practical factors.

The communication engineer may select the type of telephone set best suited to the situation at hand, and the voltage and battery current supplied to the set can, if necessary, be chosen to control the efficiency of the transmitter. Economics and the need for standardization in a system of some magnitude must also be considered.

The measurement of the over-all acoustic performance of a telephone circuit from transmitter input to receiver output presents some problems in practical application although, currently, some consideration is being given to its possibilities.

Hitherto, the transmission design and maintenance of a telephone system have usually been judged chiefly by the performance of the electric circuits between the two ends of a telephone connection.

The most important considerations are:

1. The *attenuation* of the path over the range of frequencies in the speech band, roughly from 200 to at least 3,000 cycles.

Related is the *linearity* or load-carrying capacity of the path. In effect, this is the

change in attenuation of the path with varying powers. If there are nonlinear elements in the path or if the power applied exceeds reasonable limits, the input-output power relationship will not be proportional. Distortion of the wave form of the speech currents will result. Frequencies other than those applied to the circuit will be created to degrade the speech.

A fair amount of nonlinearity can be permitted in a circuit carrying only a single speech channel. Where unwanted frequency components can be kept below 30 db or so of the wanted currents, the situation is usually tolerable. However, in a carrier or radio circuit having more than one channel separated on a frequency basis, nonlinearity in common equipment produces intermodulation between the different channel frequencies. This results in noise or cross talk. The linearity requirements in such multiple-channel systems are considerably more severe than those of single-channel systems.

A practical measure of the linearity of a circuit is to apply, simultaneously, currents of more than one frequency of the desired level and measure the intermodulation frequency components produced.

2. The relative *phase change* of the different frequencies in the speech band. In the ideal case the phases of the different frequencies in the speech band should be such as to establish the same wave form as originally transmitted. For ordinary telephony this factor can, however, usually be disregarded.

3. The amount of unwanted current or *noise* which reaches the circuit from external sources. In its effect on the listener this is added to the room noise which he experiences while using the telephone. The absolute amount of noise can be serious but is usually less important than the noise relative to the average speech power. The character of the unwanted noise is also important in determining the severity of the disturbance.

If the unwanted sounds consist of cross talk from adjacent circuits or other sources, *privacy* is also a consideration. In the general case cross talk cannot be permitted to be intelligible to a user. This means that it must be kept to less than the order of 60 db below or one-millionth of the average speech power. The magnitude of the other noise on the circuit is a factor in determining the exact amount of cross talk permissible.

4. The *impedance* of the circuit as determined by the electrical constants of the circuit. In a uniform wire circuit, it is the *characteristic impedance*, or the impedance of such a circuit of infinite length. Of course, a radio system or other equipment which may form part of a long circuit will also have an input and output impedance determined by its components.

The influence of impedance on system performance varies. It may, in the case of a wire circuit, be related to the attenuation of the circuit. It is of prime importance in connection with the operation of repeaters in two-wire circuits and in the junction of two-wire and four-wire circuits.

It always is an important factor when circuits or equipments of different types are joined together. In joining circuits it is generally desirable that they should have the same impedance. Failure to match impedances may result in increased circuit losses by reflections. If there is not a reasonably smooth impedance match throughout the effective frequency range, there may also be amplitude-frequency distortion to impair speech quality. Echoes may also result.

30. Transmission Measurement. The evaluation of telephone transmission imposes problems that do not exist in the case of power transmission, where it is comparatively simple, with a voltmeter, ammeter, or wattmeter, to make a precise measurement at any point in the circuit under conditions of essentially constant frequency.

In the case of a telephone transmitter, both the absolute power and frequency vary from instant to instant. The average a-c power which flows from a transmitter with strong talkers is of the order of $\frac{1}{10}$ mw if integrated over a long period. With a weak talker the power may readily drop to one one-thousandth of this figure. Instantaneous peaks of power may exceed ten times these figures. Correspondingly, the frequencies may vary instantaneously throughout a range from 200 to over 3,000 cps.

With a knowledge of the characteristics of the transmitter and receiver, it suffices, in general, to treat the engineering aspects of telephone transmission by a consideration of the *relative* power rather than absolute power. This means paying attention chiefly to the attenuation of the circuits—for the frequency range of interest.

In the case of wire circuits of uniform construction, it is possible to compute the loss of the circuit from a knowledge of its electrical constants, as discussed further under

Wire Circuits. It is possible in general to determine, by estimate or measurement, the characteristics of any of the other elements that may make up a long telephone circuit.

31. Transmission Units, Db, Dbm, VU. Such computations or measurements are commonly expressed in *decibels*, which has the advantage that the over-all attenuation of a circuit made up of various links can be determined by the simple addition of the decibel losses of the component sections. In addition, there is evidence that, in its effect on the ear, relative loudness is judged roughly in proportion to the logarithm of the relative energy impressed.

While the decibel is commonly used in America, some of the European countries employ the *neper*, equal to the natural logarithm (base ϵ) of the ratio of two voltages or currents. If the power ratio is the square of the voltage or current ratio, the nepers $n = \frac{1}{2} \log_\epsilon (P_1/P_2)$. One neper is equal to 8.686 db. This unit has a natural significance in connection with the mathematical treatment of transmission problems. From a practical standpoint, the decibel has somewhat greater usefulness.

It cannot be too strongly emphasized that the decibel measurement is, of itself, merely a *power ratio*. It implies nothing as to the absolute amount of power, voltage, or current that may exist unless this is definitely indicated with the use of the term.

A *transmission-measuring set* is often used, and directly calibrated, by assuming that at the input end of the circuit or equipment concerned, a steady source of 1-mw a.c. will be applied. Such devices are employed to a considerable extent in the telephone plant to check the attenuation of circuits, often at 1,000 cps. A transmission-measuring set is essentially an amplifier-rectifier, connected to a d-c milliammeter, which is correct-reading when connected to circuits of a standardized impedance, commonly 600 ohms, for trunk circuits. With other impedances, corrections must be made. As noted, the transmission-measuring set does not indicate the absolute value of the speech power that may exist in the circuit when in use. However, its reading on steady currents is, in fact, decibels above or below 1 mw, generally abbreviated *dbm*.

Another instrument, the volume indicator, or VU meter, described in Chaps. 13 and 21, is employed where a rough indication of the absolute value of the speech or music power or volume in the circuit may be desired. This consists of a meter of standardized dynamic characteristics whose fluctuations follow, roughly, the major fluctuations of the speech energy in the circuit. Its constants were originally determined experimentally to reflect the instantaneous speech-energy characteristics which are of significance in the case of some types of equipment, such as amplifiers, subject to peak overloading. It is necessary to follow a prescribed technique for reading a VU meter; the maximum excursions of the needle of the instrument are noted within a given period of time. With experience, the reading can be checked, to a fair degree, by different observers.

32. Phase and Delay Considerations. It was assumed for many years, following the early experimental work of Helmholtz, that, while the ear is fairly critical of amplitude-frequency distortion, it does not appreciate phase or delay differences between the different frequency components. In more recent years more careful experiments have revealed that this is not strictly true even for ordinary speech, and particularly is not so in the case of sounds from musical instruments. In the latter case, for highest-fidelity rendition, close phase control, as well as very-wide-band transmission, is desirable—over a range from as low as 20 cps to frequencies well above 10,000 cps, depending on the instrument.

Television transmission is also very sensitive to phase differences throughout the several-megacycle-wide frequency band involved. Telephone or data-transmission systems involving frequency modulation or pulse transmission—particularly multichannel systems—are unusually sensitive to phase distortion. Instruments are available for measuring phase delay.

Phase distortion occurs to only a slight extent in simple uniform wire-transmission circuits, either cable or open-wire, where there is no abrupt transition in attenuation as the frequency is raised. However, in *loaded-line circuits, i.e.*, circuits in which inductance has been added at intervals in series with the wires in order to reduce the

attenuation, resulting in a comparatively sharp cutoff in the attenuation-frequency characteristic, there is considerable phase change in the region approaching the cutoff frequency. Similarly, in types of equipment where the band of frequencies involved is restricted by network or filter circuits, there is also likely to be considerable phase change in the neighborhood of the transmission characteristic where the losses increase rapidly.

The effects of such phase changes can be reduced by adding to the circuit special correcting networks called *phase equalizers*. In some cases, it may be feasible to restrict the employed range of frequencies to a band which avoids the frequency range where the phase change is most severe.

The situation seldom, if ever, arises as a practical factor for ordinary speech and is scarcely noticeable except on the longest circuits. Even there, it impairs the naturalness only slightly while not seriously affecting the intelligibility.

Absolute time delay in transmission is occasionally a factor. On open-wire-line circuits the velocity of transmission approaches that of light. Nonloaded cable circuits, depending on the gage, have a velocity only a fraction of this in the voice range but approaching 150,000 miles per sec at high frequencies. The application of loading to circuits may slow down the velocity to only one-tenth this figure or less. Coaxial circuits have a high velocity of transmission, and, of course, the radio medium is high velocity.

However, delay frequently occurs in substantial magnitude in the terminal equipment of radio or carrier systems when the currents pass through networks or filters. Frequently such delay is greater than that which exists in the medium itself.

Delay, if permitted to exceed $\frac{1}{10}$ or $\frac{2}{10}$ sec, may be annoying to a user because, when he speaks into the transmitter, he receives side tone not only directly from within the telephone set itself or through the air but, also, to some extent, from reflections along the line circuit and, in particular, from the far terminal of the line which may be several hundreds or thousands of miles distant. The latter side tone, while not necessarily approaching the volume of the speech received on the circuit, may still be very annoying to the talker if it is delayed appreciably beyond the speech itself. In effect, it becomes an *echo*.

In the case of circuits exceeding 1,500 miles or so, arrangements are frequently made to introduce *echo suppressors*. These are voice-operated devices which, when speech passes through them in one direction, interrupt the transmission path in the other direction. This is feasible since the majority of longer-distance circuits are so-called *four-wire* two-path circuits, or perhaps *equivalent four-wire* where different frequencies are employed for transmission in each direction in carrier or radio systems.

If the delay in transmission from end to end of a long circuit is extremely great, involving a substantial fraction of a second, this may bring about an undesirable situation because it takes so long for the person at the receiving end to get the message and, if he wishes, to interrupt or reply to the first speaker. This can happen, particularly in cases involving a combination of unusually long land and overseas transmission paths. There is not much that can be done to alleviate this situation except to keep out of the longer circuits those elements which reduce speed, such as loading. As a matter of fact, loading is currently being installed chiefly only on intra-area exchange trunks. Some medium-distance loaded intercity circuits, installed many years ago, are being rapidly unloaded for the application of carrier systems.

33. Noise. It is evident that, while, theoretically, circuits might be permitted to extend to great distances with the expectation that the attenuation could be overcome by applying an amplifier at the end to offset it, if there are sources of unwanted current or noise somewhere along the circuit, there is a definite limit to the amount of amplification which can be employed. Otherwise the noise current may compare with, or even exceed, the wanted speech currents.

Practical experience and tests indicate that a person is disturbed by some types of noise when it is stronger than 30 db below the average energy of the speech. Actually, of course, noise greater than this can be tolerated in speech circuits, at least for short

intervals. By concentration, not too much intelligibility is lost when there is a signal-to-noise ratio as low as 10 db. Such conditions are, however, not proper standards for satisfactory day-by-day telephone service. Where telephone circuits are used for some forms of data transmission, even more severe noise limits must be met.

The character of the noise plays a part in determining the disturbing effect. In general, steady regular wave-form disturbances are less annoying than erratic or sharp high-peak-value transients. Noise which sounds like speech, even though not intelligible, is sometimes very disturbing.

Noise can be controlled in various ways—sometimes by suppressing the source of the disturbance. For example, noise which results from the induction from power lines into telephone circuits is generally not the result of the 60-cycle power field but of higher harmonics which are inadvertently produced in some generators and power-converting devices, such as rectifiers. These harmonics may involve frequencies throughout the whole voice range and, while not representing a large percentage of the power transmitted, are frequently of sufficient strength to induce substantial currents in telephone circuits, particularly open-wire circuits.

It is possible, by improving the design of the power generators or converting devices, in some cases applying chokes or filters in the power circuits, to effect considerable improvement in particular situations. By applying transpositions in the power circuits as well as by means of transpositions in the telephone-wire circuits, a substantial reduction in induced noise can also be effected. Carrier systems, as in radio, are sometimes subject to interference from static, also from radio stations having overlapping frequency ranges.

Cable circuits, by virtue of the close spacing and continuous twisting of the conductors forming a pair and also because of the metal sheath, are much less sensitive to picking up noise from external fields, such as power and other sources.

34. Balance. Most metallic circuits in good repair are inherently fairly well balanced to ground by virtue of their constructional symmetry, but if they are connected to equipment which is unsymmetrical to ground and which has any appreciable degree of *resistance* or *capacitance unbalance*, this will tend to convert cross talk or noise currents which may be picked up in the longitudinal circuit of the pair and transfer it into the metallic circuit. It is important therefore that equipment intended to be connected to balanced line circuits be carefully designed.

A common method of expressing unbalance in circuits and equipment is to measure the microamperes induced in the metallic circuit per volt applied to the longitudinal line circuit. Well-constructed and maintained open-wire or cable pairs may, of themselves, be unbalanced to the extent of 0.5 to 3.0 μa per volt when terminated in essentially perfectly balanced receiving equipment. It is desirable to have the equipment balance better than this figure. The over-all unbalance of the circuit and equipment should be a minimum in order to minimize the noise received, particularly where the noise pickup from external sources in the longitudinal circuit is substantial, as it may well be, even in shielded-cable circuits.

35. Resistance Noise. Actually, if circuits could be perfectly shielded, electrostatically and electromagnetically, from external noise of all sources, there would still remain inherent in the circuit itself a noise source which is the result of the continuous random movement of the electrons in the conductors. This so-called *resistance noise* or thermal noise is a function of the temperature of the conductor and the band width of the circuit. Its frequency distribution is essentially uniform. It is seldom a practical factor in the design and layout of most telephone circuits since it is much lower in power than the noise which usually appears from external sources or is picked up incidentally in the equipment.

The resistance noise which appears in a telephone circuit at 70°F is of the order of 140 db below 1 mw. However, there are a few practical cases where noise from external sources is low, such as in a long overseas submarine cable, where resistance noise may be a controlling factor.

After all ordinary measures are taken to reduce unwanted currents or noise which

enters a telephone circuit, the only remedy possible is to ensure that the speech level does not sink so low in traversing a circuit that it approaches the level of the noise currents. This can be accomplished by applying amplifiers (*repeaters*) to raise the level of the speech currents at frequent intervals along the circuit. In laying out a circuit and spacing the repeaters, consideration must be given to the number which will appear in the over-all circuit, and whether the particular circuit involved is only one of several which may be connected in tandem. Even though a certain signal-to-noise ratio may be permissible for a single link circuit, if several may be connected, each must be designed to have a somewhat higher signal-to-noise ratio than desired in the over-all multilink circuit.

Noise is commonly measured by connecting to the circuit a so-called *noise-measuring set*. This has some elements in common with a VU meter, except that it is usually more sensitive and its frequency-transmission characteristics are adjustable to simulate the telephone receiver plus the approximate human-ear sensitivity to disturbances. VU meters have sometimes been used with the necessary extra amplification in the absence of a regular noise meter. Noise measurements are commonly expressed in decibels above a base, which is approximately at 85 db below 1 mw.

36. Impedance. As noted previously, the impedance of circuits and equipment which make up long telephone circuits must be carefully considered. Failure to match impedances, where two circuits or equipment units join, may result in increased losses and/or transmission-frequency distortion. For example, the characteristic impedance of an open-wire circuit, having 12-in. spaced conductors, is of the order of 600 ohms throughout both the voice and carrier frequency range. A nonloaded cable pair has an impedance of the order of 100 ohms in the carrier range. If two circuits of these respective types and of substantial lengths are directly connected together, the over-all loss will include not only the attenuation loss of the individual circuits concerned but a reflection loss of about 3 db, which may be serious.

The remedy for impedance problems is either to design the circuits and equipment so that they conform to uniform impedance values, 600 ohms being an accepted standard for most regular telephone plants, or, if junctions are to be made between circuits having different impedance, to include transformers, generally called *repeating coils*, having proper turn ratios, at the junctions in order to minimize reflection effects.

Sometimes impedance-transforming networks or autotransformers can be applied, particularly where it is desired to preserve the d-c continuity of the circuit. The treatment is not always simple because the impedance of line circuits and devices is not constant with frequency and includes the reactive as well as resistive components.

37. Transmission Standards. With telephone sets of the general type described in Sec. 2, it has been established, by measurements and tests, that the requirements for reasonably satisfactory transmission are satisfied if the over-all attenuation between two users is not greater than about 25 db. It is assumed in this connection that the transmission characteristics of the over-all circuit provide a band width from roughly 200 to over 3,000 cps, at which points the attenuation is no greater than 10 db above the mid-band attenuation, and that the variations throughout the band do not exceed a few decibels. It is assumed, also, that the circuit noise entering the receiving telephone set is at least of the order of 40 to 50 db below the speech power leaving the transmitting telephone set, with the cross talk at least 60 db below.

The transmission designer will need to consider many factors in specifying the design objectives for the different trunk and loop circuits involved, obviously with a detailed knowledge of the economics of the situation. In a nationwide system, in addition to two loops at the terminals of the connection, there may be other trunks between the end offices, and perhaps also between *toll centers*, *primary centers*, *sectional centers*, and *regional centers*, these being the successive steps in an echelon of switching organization. In fact, a complete intercity connection between subscribers may have as many as eight or nine links. With automatic switching, there is no serious barrier from the standpoint of speed of service in having many links, and considerable advantage in flexibility. Manual switching previously placed a greater premium on direct circuits.

The total permissible loss for each link is not necessarily a proportionate fraction of the 25 db suggested as an over-all design objective nor is 25 db necessarily a limit for all the future because, with time, improvements in equipment and techniques will probably permit continuing the trend in the direction of still lower-loss transmission standards.

In general, it costs very little more to provide close-to-zero-loss long-distance trunks, where amplifiers are needed in the circuit in any case. It is, of course, also desirable to engineer the circuits in such fashion that different connections are reasonably alike in volume irrespective of the distance involved.

The use of volume indicators, or VU meters, has not, in general, been very extensive in telephone-system practice except in the maintenance and setting up of special circuits, such as broadcasting, where the order of magnitude of the power cannot always be predetermined as well as for regular telephone use.

However, they are being tried out for periodic checkups of over-all system transmission performance by connecting them across a point in the circuit at the office end of a user's loop and, over a period of time, noting the VU readings when speech is being received from a distant user. A better alternative might be to connect them across the user's set itself, but this is not very practical.

Such readings at the central office are significant as to the over-all transmission a user is receiving. They include the effect of the complete transmission-line circuit from the distant user and, in addition, of course, involve the loudness with which the distant user is speaking and how he holds the telephone. Some skill is needed in interpreting the results of such a volume-indicator survey and considerable data are necessary before judging as to what, if anything, may be wrong with the telephone circuits or equipment.

Based on subjective studies, it appears that a received volume above −34 VU or so would be generally rated by the receiving party as good, between −34 and −40 VU as fair, and below −40 VU as poor. These figures assume a receiving loop with attenuation of about 5 db and a 300-type telephone set somewhat older in design than the 500 set described in Sec. 2. Of course, as noted previously, a low measurement in any particular case may be due either to faulty circuits or equally as well to an habitually weak talker or one having other imperfect telephone habits.

WIRE CIRCUITS

38. Types of Wire Circuits. The first circuits for telephoning any distance followed earlier telegraph practice and consisted of a single bare wire supported by insulators on poles, with a ground connection at each end. It was soon found that, when two such *ground-return* circuits were operated on the same pole line for a substantial distance, induction between them caused considerable speech on one circuit to be overheard on the other. A great reduction in this *cross talk* resulted from the general adoption of a *two-wire circuit, i.e.*, with metallic rather than a ground return.

Circuits are subject to interference not only from other telephone circuits but from nearby power lines as well as static and lightning impulses. The two-wire circuit is, of course, not, of itself, adequate to control the noise situation in all cases, as discussed in Sec. 33. Unwanted currents may, of course, also arise inadvertently in the equipment, from power sources in amplifiers and the like.

Wire circuits employed for telephony range from the simple twisted pair through various types of multipair cable circuits and open-wire lines to the more complex coaxial conductor. The variations in size and performance are numerous.

Simple insulated pairs may be used for interior wiring to telephone sets within a building. In a typical case use is made of one or two pairs of 20-gage copper-steel conductors, with plastic insulation, and a jacket colored to harmonize with room trim and thus remain inconspicuous where it is not possible to conceal the wire. In large-scale installations, cables of many pairs, usually 24-gage, may be run within the walls, in ducts or runways.

A much more substantial pair construction is used to provide the *drop wire* from a

telephone pole to a user's home. Since spans of 200 ft or more are sometimes necessary, the tensile strength of such wire is important. Eighteen (18) gage is usually employed, the wire being of a high-strength alloy or possibly copper-clad steel. Such wire has a strong insulation, such as buna rubber, and is jacketed with tough abrasion-resistant neoprene or other plastic to afford protection when it comes in contact with tree branches.

The transmission characteristics of interior wire or drop wire are not a prime consideration since the lengths involved are short.

39. Loop Circuits. For transmission over the loop circuits from users to central office, a cable, made up of many pairs of wires under a moistureproof sheath, is most commonly employed. This cable may be laid underground, in ducts in cities, or perhaps supported aerially from poles. Where single cables are involved, there is an increasing use of direct burial underground without ducts.

Depending on the demands of service and the distances involved, such cables may range from a few pairs upward to several thousand. The gages range from No. 19 to No. 26 B. and S. The increasing cost of copper has encouraged the application of aluminum wire in some instances.

Older practice has been to insulate the pairs making up the cable with spiraled paper or paper pulp. Currently, plastic material, such as *polyethylene*, is being used for insulation. Also, older construction generally employed thin lead sheaths over the bundle of conductors. Current practice is to use a plastic, such as polyethylene or polyvinyl chloride, sheath with an aluminum or a steel and aluminum sheath underneath.

Sheath damage, with consequent leakage of moisture into the cable structure, presents a serious problem, particularly in the case of paper-insulated conductors, where the moisture penetration may affect the transmission on many pairs. This has led to fairly extensive pressurizing of the cables so that a leak of substantial size can be detected rapidly and moisture can be prevented from entering the cable as long as the pressure is sustained.

The pressure on the interior of the cable is applied, usually by means of nitrogen gas in cylinders or air from compressing and drying equipment, at about 10 psi. Facilities for measuring the pressure at various points along the cable make it possible, by comparing measurements, to locate a leak to a fair degree of accuracy.

40. Open-wire Line. Where not too many circuits are involved, open-wire lines are still employed to some extent in loop circuits. This is the case in some rural areas. Open-wire lines may have *crossarms* supporting from 2 to 16 glass insulators. There may be from one to four crossarms, sometimes more, totaling 20 or more pairs of wires.

The two wires constituting a pair are commonly spaced 10 or 12 in. apart, with the same spacing between pairs. On some lines, where it is desired to use carrier systems with frequencies above the voice range, it has been found useful to employ pair spacings as close as 8 in. to minimize cross talk. Experiments are also currently being made with 6-in.-spaced pairs, where one or both wires have a thin cover of plastic insulation to minimize the chance of the conductors touching when swinging in the wind.

The sizes of open-wire conductors range from 80 to 165 mils (diameter). In the case of short lines, where the attenuation may not be so critical, galvanized-steel wire is sometimes employed. This permits greater pole spacings, up to 600 ft, because of the additional strength. In addition, there is the lower cost of the conductor itself. Copper-clad steel has also been employed. Galvanized-steel wire and, to a lesser extent, copper steel have the disadvantage of having higher resistance and, consequently, greater attenuation than copper of approximately the same size at high frequencies; however, copper-clad steel is essentially as good as all-copper wire.

Transposition. Cross talk is a factor whenever circuits exist in close proximity in either cable or open-wire construction. In the case of open-wire lines, cross talk is minimized by interchanging the positions of the wires of a pair at intervals to form a

so-called *transposition point.* Where there are a number of pairs on an open-wire line, the location of these transposition points differs for the different pairs and must form an orderly pattern. The transpositions for the different pairs are generally chosen from fundamental types shown in Fig. 6 of Chap. 27. In the case of pairs used for high frequencies, the locations of the transpositions must be quite precise.

In the case of pairs of wires in cables, the inductive effect between pairs is minimized by continuously twisting the different pairs. It is arranged, in laying up the cable, that those pairs which are in closest proximity have different rates of twist. A cable pair may have any one of several different rates of twist, depending on the location of the pair in the cross section of the cable.

41. Trunk Circuits. Intracity trunk circuits are almost exclusively of cable construction, either underground or overhead. They are frequently loaded at intervals of from 3,000 to 6,000 ft, as described later. Intercity construction may also have somewhat similar characteristics except that the gage employed, mostly No. 19 gage, is larger than the pairs generally employed in loop circuits.

In intercity cables, the pairs are sometimes used as voice-frequency *four-wire circuits, i.e.*, for one-directional transmission. All the pairs for a given direction of transmission are often grouped and separated from those of the opposite directional transmission. In a few cases these groups of conductors have been shielded with a covering of thin foil to minimize further the cross talk between the opposite directional groups, particularly when the pairs carry high frequencies. In most cases of long carrier systems having the same frequencies in opposite directions, separate cables are used.

Cross-talk problems are more severe between opposite directional circuits than where the circuits concerned transmit in the same direction. This is because the difference in levels of the speech currents near the input or output of repeaters in the opposite directional pairs may be substantial. In general, intercity cables are more carefully controlled in uniformity of construction, particularly to minimize cross talk, than exchange-area cables.

Where *phantom circuits* are involved, the pairs are first grouped, two pairs together, termed a *quad*, and these are given a separate spiraling in addition to the spiraling of the individual pairs in order to reduce quad-to-quad cross talk as well as quad-to-side-circuit cross talk. There is also a *star-quad* nonphantom construction, not so common in America but more common in Europe, in which the two pairs making up the quad have their two respective conductors arranged noninductively at the diagonal corners of a square. Thus the side circuits are not individually spiraled but the star or *spiral-four* quad is spiraled as a unit. Such a quad saves some space in the cable.

Quite frequently long-distance intercity cables and distribution cables are laid directly underground without being drawn into underground ducts. Armor wires or tapes for physical protection and perhaps some extra defense against lightning are added over the cable. Some of it may be laid by tractor-drawn plow-type machinery which does not require that a trench be opened up and backfilled.

For some intercity lines, there is still a fair use of open-wire construction, particularly where not too many circuits are required. The cost per circuit mile may be lower, among other things, because the attenuation is lower and they require fewer repeaters than cable circuits. These open-wire lines are not dissimilar to those used in loop circuits but usually are more rugged, having heavier poles, shorter spans, 150 ft or less, and larger wire sizes, in some cases 165 mils. Such well-constructed open-wire lines have a good degree of reliability, although in areas where there is much sleet, they are sometimes vulnerable.

42. Coaxial Cable. The *coaxial* conductor is a preferred form of construction where the transmission of exceptionally high frequencies is involved. It consists essentially of a flexible cylinder surrounding an interior solid or stranded conductor. From an attenuation standpoint, it is good for high frequencies for, in the case of the outside conductor, the high-frequency resistance is especially low because the *skin effect* is minimized. Furthermore, a coaxial structure tends to be self-shielding at the high frequencies and is not vulnerable to external fields. If the insulation is largely air, as

is the case for the coaxial circuits used in the telephone plant, the dielectric losses can be kept low and this again promotes its effectiveness for frequencies in the megacycle range. In this case the central conductor is supported by thin, circular-shaped spacers of polyethylene so that, effectively, the dielectric is about 95 per cent air.

A standard coaxial conductor, generally used for carrier and television transmission, has an internal diameter of the external pipe of 0.375 in. The inside conductor consists of a solid wire. The outside conductor is a long thin copper strip which has been folded in cylindrical shape, with a butt joint, the joined edges having small intermeshed teeth, and the whole bound together tightly by a serving of thin steel tape. Great precision is required in manufacture to maintain uniformity of transmission and impedance characteristics. Its flexibility permits it to be made up in cables which may contain several coaxials as well as numerous small-gage pairs, if desired. Eight coaxials, as well as some ordinary pairs, are often included in such a structure, as shown in Fig. 18.

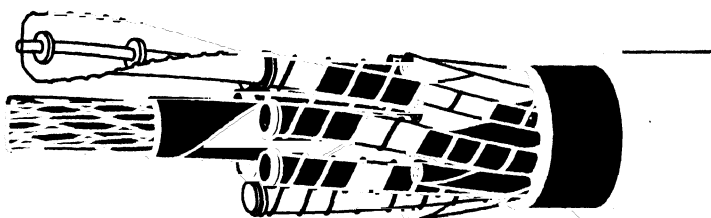



FIG. 18. An assembly of eight coaxial cables.

Solid dielectric coaxial lines are often employed for short distances in connection with the wiring of high-frequency equipment. In general, their losses are too high to make them economical for the longer circuits except where great compressional resistance is required, as in a deep-sea cable.

43. Local Video Lines. Another form of special conductor has been employed, to some extent, in trunk circuits, around cities, for television transmission. This is a pair rather than a coaxial. A coaxial circuit, while well shielded at high frequencies, is not well shielded at very low frequencies. For the transmission of television programs on a noncarrier basis, where frequencies approaching zero are involved, a balanced pair can be made to be insensitive to low-frequency, as well as high-frequency, interference.

One form of this special high-frequency *video pair* consists of two No. 16 gage wires, each of which has a thin cover of gas-expanded polyethylene, and which are twisted together with two interstice fillers of expanded polyethylene over Fiberglas string. It is covered with a polyethylene wrap and a longitudinal copper tape shield.

It has an attenuation of something like twice that of standard 0.375-in. coaxial conductor. If used for ordinary television transmission for distances of greater than 4.5 miles, it must be equipped with repeaters at 2- to 3-mile intervals. Frequencies up to 10 Mc can, if necessary, be transmitted over such conductors.

44. Wire-transmission Characteristics. In a wire circuit of uniform construction, the attenuation in *decibels* per unit length can be computed from the formula

$$A = 6.141 \sqrt{\sqrt{(R^2 + \omega^2L^2)(G^2 + \omega^2C^2)} + RG - \omega^2LC}$$

where A = db per unit length
R = resistance per unit length
L = inductance per unit length
G = leakage per unit length
C = capacitance per unit length
ω = $2\pi \times$ the frequency in cps

Of course, the same unit length must be employed in the computation, and the units expressed consistently as, for example, ohms, henrys, mhos, and farads.

For very low frequencies, the computation poses little difficulty since the constants involved are essentially d-c values. However, as the frequency is raised, the so-called constants take on other values because of skin effect and other factors.

Skin effect and related phenomena result from the fact that the currents at higher frequencies tend to redistribute themselves throughout the cross sections of the conductors and travel a path of least impedance. Thus, in the case of a simple conductor

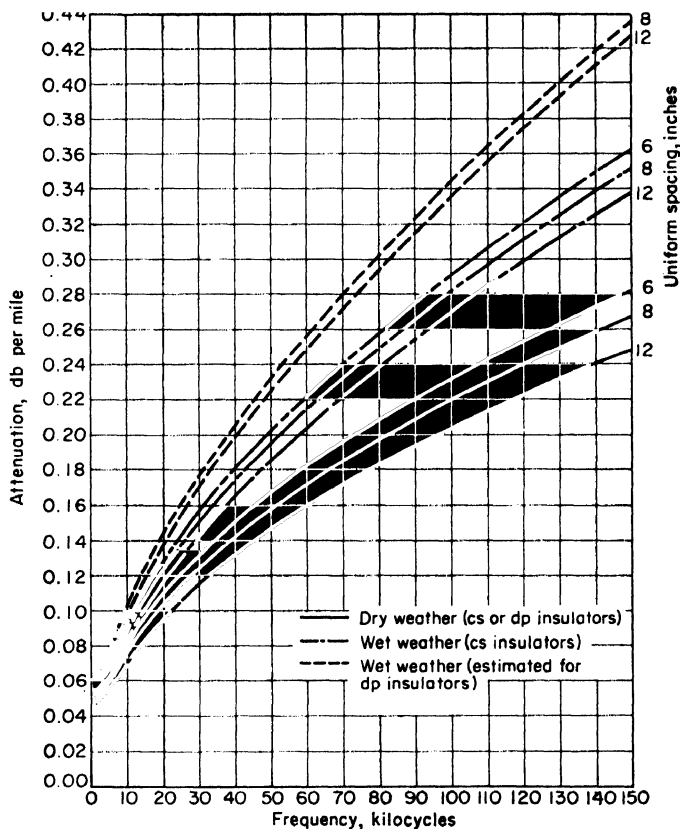



FIG. 19. Attenuation of 128-mil copper. Calculated for hard-drawn copper at 68°F with 53 pairs of insulators per mile.

of circular cross section relatively remote from the return circuit, the current will tend to avoid the center of the conductor and crowd the outside layers to a degree, depending on the frequency. In the case of very high frequencies, the current effectively travels only in a relatively thin surface layer. Where the conductors of a two-wire circuit are in relatively close proximity, as in a cable pair, this same tendency results in the currents' crowding the portions of the conductors which are closest to each other.

It is evident that this redistribution of the currents as the frequency is varied will not only determine the effective resistance of the conductor but will also affect the inductance and capacitance per unit length.

45. Line Leakage. Another difficulty in predicting the attenuation is the matter of the effective leakage. For higher frequencies, this not only is made up of the d-c

component but also includes effectively all dielectric losses in the circuit. For higher frequencies, with solid dielectrics, as in cables, these losses can be appreciable. Together with the resistance of the conductor, they may also vary with temperature and time and other external circumstances.

46. Environment Factors. Open-wire lines are unusually susceptive to variations with time. The chief offender is the variation of leakage with degree of moisture on the insulators. Not only may this leakage result from direct surface conduction over the insulator when coated by rain or fog, but the dielectric losses in the insulators may play an important part when the frequencies are high. An unusually serious condition exists when the insulators and wires are covered with ice in some form, with sleet or perhaps hoarfrost, a crystalline deposit which, in some special conditions of fog and temperature, may coat the wires and insulators to a depth of several inches. These conditions, with ice having an extremely high dielectric constant, may completely upset the attenuation, as will be appreciated from some of the data included in curves given in Figs. 19 to 23.

Fig. 20. Attenuation due to ice, snow, and frost on open-wire pairs of hard-drawn copper.

In general, the characteristic impedance of the wire circuit is less affected by the variations in line constants with time and temperature.

Typical characteristic impedance curves are given in Fig. 24 for 12-in. spaced open-wire conductors for frequencies from 0 to 150 kc, also in Fig. 25 for typical cable circuits from 0 to 150 kc for No. 19 gage conductors.

47. Loading. It is generally possible, with most common wire circuits, to reduce the attenuation by increasing the inductance per unit length. This is termed *loading.* It can be carried out by uniformly surrounding the conductor with a thin layer of magnetic material of high permeability, which is known as *continuous loading.* This has been used, to some extent, in the construction of some long submarine-cable circuits for telegraphy as well as telephony.

However, loading is usually introduced in telephone circuits by connecting *loading coils* in series with the conductors at intervals. These consist of a magnetic core having two windings, through one of which passes one conductor and through the other the second conductor of the pair. The spacing of the loading coils must be a fraction of the wavelength involved for the highest frequency it is intended to transmit. Such *coil loading* has a *cutoff frequency* beyond which the attenuation rises to extremely high magnitudes. Within the transmission band the attenuation is usually of the order of one-third to one-half that of the corresponding nonloaded pair.

FIG. 21. Copper-clad-steel pairs of 128-mil wire, 68°F and 53 pairs of insulators per mile.

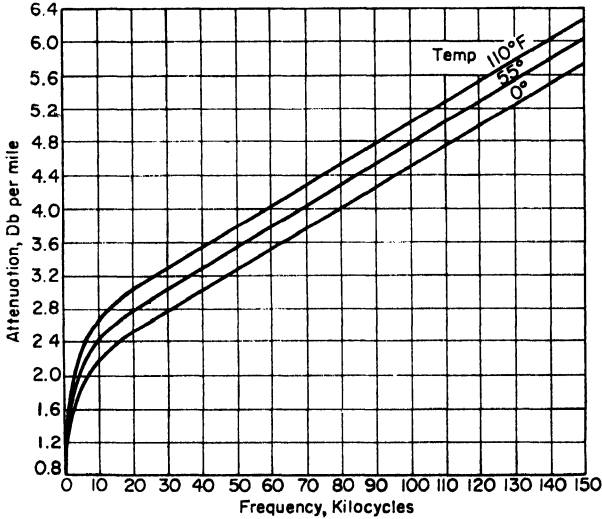



FIG. 22. Typical attenuation of 19-gage cable pairs.

The addition of loading raises the characteristic impedance of the circuit, and this tends to exaggerate the leakage losses. Loading is therefore not generally regarded as a good expedient where it is expected to use the highest frequencies on a line, where leakage losses may predominate.

In the earlier days of telephony before the advent of vacuum-tube amplifiers and the use of high frequencies to transmit several channels over the same conductors,

loading was common on the longer wire circuits, both open-wire and cable. These practices have been largely discontinued, and, indeed, many circuits, originally loaded, have been unloaded.

Loading is currently being installed in the telephone plant mostly on the relatively short intracity cable circuits where, alone or in combination with repeaters, it is the most economical method for obtaining satisfactory overall losses. Carrier techniques are becoming cheaper, and with the prospect of a more extensive application of transistors, it seems not unlikely that loaded voice usage may be limited to circuits up to 8 or 10 miles in length in the not too distant future.

48. Phantom Circuits. Two sets of conductors, each individually used for two-wire or four-wire transmission, can be connected, by the use of *repeating coils* or phantom coils, to provide a third so-called phantom circuit, as shown in Fig. 26. The currents of the two side circuits flow in opposite directions for the pair, while the poling of the windings of the repeating coils is such that the phantom-circuit currents flow in the one direction over both sides of one pair and in the opposite direction over both sides of the other pair. Therefore, there is little or no cross talk between the phantom circuit and either of the two side circuits if the repeating coils are accurately wound and tapped and the line is well balanced.

Fig. 23. Attenuation of high-frequency conductors.

Fig. 24. Dry-weather impedance, 12-in. copper pairs, 68°F and 53 pairs of insulators per mile.

A 50 per cent increase in the circuits provided by a given number of conductors is thus obtained for the cost of the repeating coils, which must be installed at each terminal, with some complication in the arrangements needed to minimize the cross talk between not only the phantom and side circuits of a quad, but also the phantoms and side circuits of other quads in the same structure.

The attenuation of such a phantom circuit, because of the increased amount of

copper involved, is somewhat less than that of the side-circuit pairs. Its characteristic impedance is also slightly lower. Phantom circuits have been used, to a fair extent, in the past in the trunks of the telephone system. However, their use is uneconomical in most cases now.

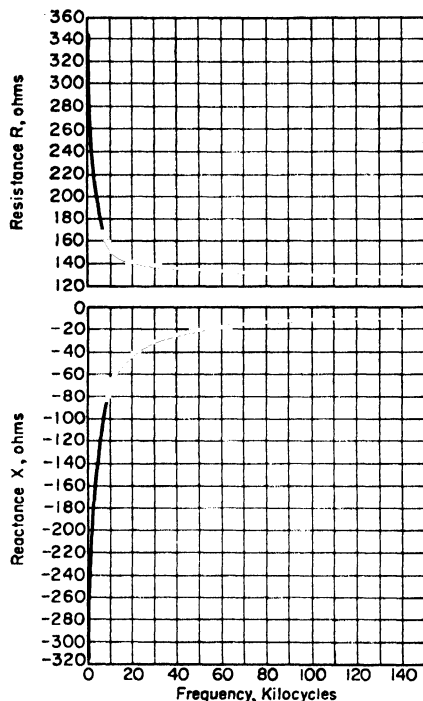



Fig. 25. Typical values of characteristic impedance, 19-gage pairs, 55°F.

49. Composite Sets. When a wire circuit is used for telephony, it commonly requires the transmission of frequencies no lower than about 200 cycles. This leaves a range from this frequency down to zero available for other purposes, such as d-c signaling or telegraphy. A separation between the frequency ranges is made by a network of inductances and capacitances, known as a *composite set.*

A simple composite-set circuit is shown in Fig. 27. The inductances are of the order of 3 henrys and the capacitances 2 μf. The ground connection makes it possible to employ the two sides of the pair in ground-return circuits if it is desired to take advantage of the two separate paths provided thereby. Such composite-set arrangements are employed chiefly in trunk circuits.

REPEATERS

Repeaters are commonly installed at intervals in either wire or radio systems. The output of a repeater should, in general, be a faithful reproduction of what enters the repeater, except that the magnitude is stepped up, depending on circumstances, from a few decibels up to as much as 100. This is referred to as the *gain* of the repeater.

50. Four-wire Repeater. The simplest telephone repeater consists of two separate amplifiers connected in the two sides of a four-wire circuit to amplify in opposite directions, as in Fig. 28. At the input of the amplifiers, such a *four-wire repeater* may include transmission *equalizers* which provide amplitude-frequency distortion opposite to that which occurs over the line circuit which the telephone currents have traversed. The equalizer generally offers relatively high loss for the lower frequencies and low loss for the higher frequencies. The net effect is to ensure that the frequency-amplitude distribution of the speech power is

Fig. 26. Arrangement of phantom circuit.

essentially the same at the output of the repeater as it was at the sending end of the circuit.

51. Two-wire Repeater. A more complicated situation arises in the case of a two-wire circuit used for speech in both directions. Here, if two amplifiers were directly connected to the line circuit in each direction to form a repeater, the gain of the amplifiers would considerably exceed the losses in the loop circuit thus established and the system

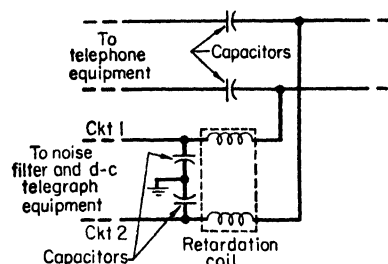



Fig. 27. Composite set circuit.

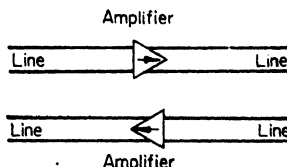



Fig. 28. Four-wire repeater.

would *sing*. This led, at an early date, to the adoption of the two-wire two-directional (22-type) repeater.

Here, instead of connecting the respective input of one amplifier and output of the other amplifier directly to a line circuit, there is introduced a balancing coil or so-called *hybrid coil*, as shown in Fig. 29. The output of one amplifier is connected to and passes through symmetrical but opposite poled windings on the coil, one path leading to the line circuit and another to a line *balancing network* whose impedance simulates that of the line. The input to the other directional amplifier is connected to another winding on the same core whose impedance is arranged to match the input of the amplifier. If there is a good impedance match between the line and network circuits, there will be a relatively high-loss feedback path around the two amplifiers. This permits substantial gain to be provided for each direction of transmission without singing.

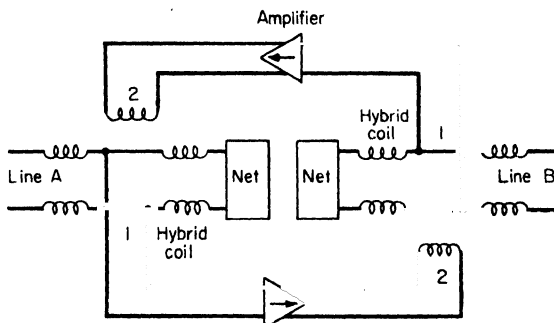
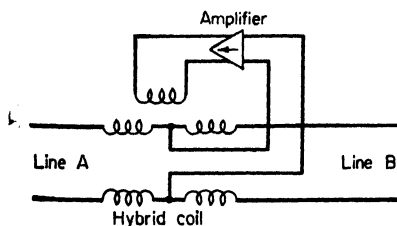



Fig. 29. 22-type repeater.

In a practical case, of course, the balancing network does not provide a perfect simulation of the line impedance, particularly if the line is short electrically. The 22-type repeater gains in each direction are usually limited to something of the order of 20 db, often less than this when there are many repeaters in tandem on the circuit.

Fig. 30. 21-type repeater.

52. 21-type Repeater. A circuit where the conductors are the same in each direction from the repeater point lends itself to the operation of a so-called 21-*type repeater* in which the two directions of transmission pass through one amplifier circuit, as shown in Fig. 30. Such 21-type repeaters are not common and are limited to occasional use where only one repeating point is necessary. The ability to balance each side of the circuit independ-

ently usually makes it possible to obtain higher gains by the use of 22-type repeaters than with 21-type repeaters.

In the case of either the 22-type or 21-type repeater, equalizers may also be needed in order to control the transmission-frequency characteristics of the over-all circuit as noted previously.

53. Negative-impedance Repeater. A special type of two-wire repeater, employing only a single amplifier, has also come into extensive use, particularly for the shorter circuits, in recent years. This is known as the *negative-impedance repeater*. A schematic circuit of a *series* negative-impedance repeater is shown in Fig. 31. Some relationship to the 21-type repeater is evident, and it will be noted that the metallic circuit through the repeater is continuous, which is a matter of some importance where it is used in connection with d-c supervisory or signaling functions. The amplifier, often termed a *converter*, which supplies the negative-impedance function, is actually a controlled-feedback circuit, the associated network being so adjusted as to supply necessary frequency-phase control in the feedback circuit in relation to the line impedances and the desired degree of amplification. Various specific amplifier circuits are possible, generally of the push-pull type.

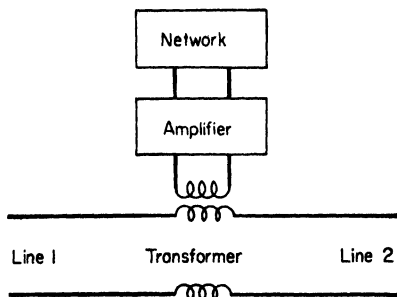



FIG. 31. E1 telephone repeater.

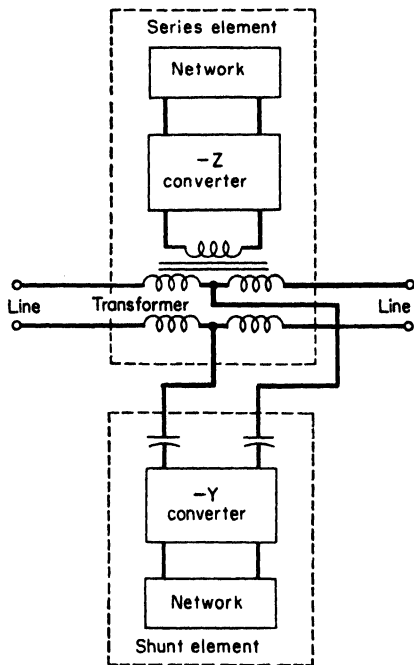



FIG. 32. Bridged-T repeater.

Shunt-connected negative-impedance repeaters are also possible, and in cases where high gains are desired, with minimum impedance irregularity, both series and shunt repeaters may be employed in combination. The schematic circuit of one such arrangement is shown in Fig. 32. It will be noted that, in effect, one converter-network circuit is in series with the line while the second one is in shunt with the line. In this case both networks must be properly adjusted to operate with particular line circuits.

Repeaters are also an essential element in the operation of carrier systems where many telephone channels are provided over the same set of conductors using a wide band of frequencies. Such repeaters must, of course, be designed to function satisfactorily over the wide band. In most cases, they require not only equalizers which ensure the desired amplitude-frequency characteristics but also some means for varying the amplification and degree of equalization with time to take care of the variations of the attenuation of the line circuits with temperature or weather. Phase equalization is sometimes necessary in such repeaters.

While, in the past, vacuum-tube amplifiers have generally been employed in repeat-

ers, it seems likely that there will be a gradual shift to the employment of transistor amplifiers, with a resultant reduction in size and power required, especially in intracity circuits where this is most important.

In a long radio circuit, particularly of the line-of-sight microwave variety, repeaters may be required at the limits of the line-of-sight distance, ranging from 25 miles upward, depending on terrain conditions. With scatter propagation, much greater distances are possible. Such repeaters can simply consist of r-f amplifiers whose terminals are connected to antennas from each direction. However, frequency translation is involved at such points.

54. Tandem Operation. The relation between attenuation, noise, and repeater spacing has already been noted, and frequently this is the only factor involved in the repeater layout of four-wire circuits.

Cross talk. However, cross talk is also a factor in some cases where it may be desired to employ substantial repeater gains. On wire circuits, where the two directions of transmission are in the same cable, cross talk could conceivably reach a point where the loss between the circuits connected to the output and the input of opposite directional repeaters would not be sufficient to prevent singing. This is not a very likely situation since cross talk between the high-level output of one circuit and the low-level input of other circuits would, no doubt, earlier have reached an undesirable point from the standpoint of privacy. The permissible gain at a repeater point must, however, be carefully considered in relation to the cross-talk situation which exists in the connecting cable or open-wire line.

In the case of repeaters on radio circuits, the cross talk between the input antenna in one direction and the corresponding output antenna in the other direction poses a somewhat similar problem. Since the desired gains in repeatered radio circuits are usually extremely high, the designer may be forced to use different opposite directional frequencies and thus supplement the directional characteristics of the antennas by frequency separation using filters. Frequencies of the two directions of transmission may, in this case, be interchanged within the repeater structure by frequency-translation equipment. Arrangements of this sort are employed in some types of carrier systems. Using different types of wave polarization also offers advantages in the case of two-directional microwave radio.

Impedance. In connection with *two-wire* repeaters, where the same frequencies are normally employed in opposite directions, the tandem operation of repeaters imposes special problems. All types of two-wire repeaters are dependent on the impedance characteristics of the wires to which they are connected. In a very long wire circuit, the impedance facing a repeater is essentially the characteristic impedance of the circuit and may not depend much on the condition of the circuit at the far end. If the line is of substantially uniform construction, it is possible to simulate the impedance of such a circuit in a balancing network quite accurately and, except for small changes with time and temperature or other weather-related circumstances, the balance provided is stable, and a fairly high repeater gain can be maintained. The gain will never, however, approach the loss of the over-all circuit to which the repeater is connected.

It may be desirable to introduce more repeaters in a long two-wire circuit at reasonable intervals. In this case, the connecting lines probably being electrically shorter, the impedance which faces the repeater will be determined by the characteristics of the line circuit plus the termination impedance, which could be a telephone set, another repeater, or other piece of equipment. If it is another repeater, the gain provided by its amplifiers may, in turn, help to reflect the impedance characteristics of a successive section of line plus some other termination or repeater equipment.

Gain. Thus, in general, the amount of gain permissible at each repeater point in a long two-wire circuit, having many repeaters, will be less than a single repeater established in a long circuit. Such tandem operation of repeaters can, however, by the necessary control of the impedance of terminal equipment and other repeater equipment, more nearly compensate for the attenuation of a long circuit than can a single

repeater installed at, say, its middle. The tandem operation of negative-impedance repeaters, particularly where both series and shunt devices are employed, can be quite effective in this connection.

However, under practical conditions, no two-wire repeatered circuit can quite approach the low over-all transmission loss of a four-wire repeater system, where the amplifiers and circuits are independent in each direction. The over-all performance of such a four-wire circuit from a transmission standpoint is generally determined by the balance provided at the hybrid coil, which is used to join the four-wire circuit to a two-wire circuit at the terminals.

CARRIER SYSTEMS

55. Carrier System Elements. The elementary carrier telephone system contains the same basic elements as a radio system. As shown in Fig. 33, the transmitters of several channels connect to a series of *modulators* fed by different high-frequency carrier-current sources, thence to selecting circuits or electrical *filters*, the outputs of which are joined to a common line circuit. At the receiving end similar selecting circuits or filters separate the modulated carrier currents of the different channels and lead to *demodulators*, by means of which the original voice currents are derived, thence to the corresponding telephone receivers of the different channels.

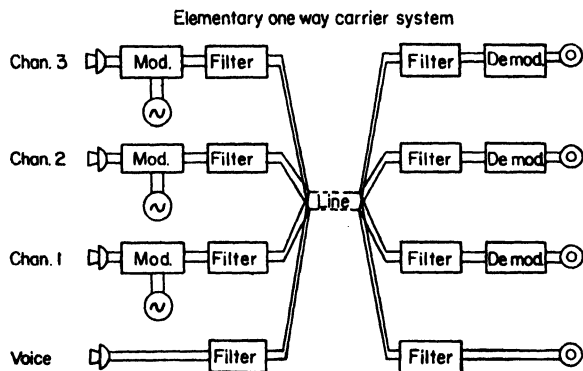



FIG. 33. Elementary one-way carrier system.

Two-way operation requires paths in the opposite directions. It is possible, theoretically, to provide two-way operation using the same carrier frequencies in opposite directions on the same set of conductors, with balancing arrangements similar to those of 22-type repeaters. However, this is not usually as practical as independent pairs of wires for transmission in opposite directions or *equivalent four-wire* operation using different frequencies over the same line circuit.

Modulation. The process of modulation may consist simply in *amplitude modulation, i.e.*, the variation of the amplitude of the carrier current in accordance with the variations of the speech wave. Actually, in this case, the resulting current, as brought out elsewhere in this handbook, consists of a mixture of the carrier current and currents of the two so-called *side bands*. The latter consist of the frequency of the carrier minus the modulating frequency of the speech at the instant, and the carrier frequency plus the modulating frequency of the speech at the instant. As has been noted, ordinary speech involves chiefly a band of about 3,000 cycles width, hence the term "side bands."

More refined modulation techniques are available by using combinations of circuitry and filters, the most common being the *single-side-band* (SSB) technique in which only one side band is transmitted, eliminating both the carrier and unnecessary complementary side band. All the intelligence is carried by the one side band, and this saves frequency space and power transmitted over the wire circuit.

Most types of carrier systems employ SSB because frequency space is expensive on wire circuits since the attenuation rises rapidly with frequency. The transmission problem is simplified thereby. However, where the distance of transmission is not too great and extreme simplicity of equipment is desired, double-side-band amplitude modulation is still employed.

Carrier systems occasionally employ other modulation techniques, such as *frequency modulation*, which requires greater frequency space than amplitude modulation. Forms of *pulse modulation*, which are, in general, even more lavish in the use of frequency space, have also been employed experimentally and, to a very limited extent, commercially.

Selectivity by simple tuned circuits has been employed to a slight extent but, more commonly, several-section *band filters* furnish the needed channel separation. For modulation and corresponding demodulation, the earlier systems employed vacuum tubes almost exclusively. Later systems employed copper-oxide diodes, while still more recently transistors or semiconductor diodes are being used in the newer carrier systems for the various functions of high-frequency generation and amplification, as well as modulation and demodulation.

56. Carrier Application. While carrier systems offer the chance to obtain more communication paths over a single set of conductors, their use must be carefully considered in each situation to determine whether it is less expensive than to provide additional separate conductors. This is because the cost of providing the equipment to carry on the necessary functions noted above tends to offset the savings in conductors.

In general, the use of carrier systems is profitable for the longer distances and has hitherto been almost entirely limited to trunk applications. However, the cost of the equipment has decreased and it seems likely that it will only be a matter of time before carrier systems are employed for distances as short as a very few miles, even for rural subscribers' lines. At the moment, the lower point of economical operation is of the order of 10 to 15 miles.

The earliest carrier systems were applied chiefly to open-wire lines where the high-frequency attenuation is comparatively low. They provided three or four additional channels with frequencies up to about 30 kc and with repeaters spaced at intervals from 100 to 150 miles. Later systems were designed for cable conductors using frequencies up to 60 kc and providing 12 channels with repeater spacings of about 17 miles. Then carrier systems using coaxial conductors came into being, providing up to 1,800 channels using frequencies up to 8 Mc and with repeater spacings of 4 to 8 miles. The latter systems can also be employed for wide-band transmission of television programs. By the end of 1956 something like 90 per cent of the circuits over 25 miles long were provided by carrier in the Bell system.

57. Power-line Carrier. Power-line carrier systems are extensively employed by power companies. In principle, they follow the general design of others using communication conductors, but the wire circuits in this case are the power conductors themselves. The connection to them is made, in general, through *coupling capacitors* capable of withstanding the high voltage of the power lines, at the same time passing the high frequencies of the carrier systems. Such carrier systems, in addition to their use for telephony, have also been employed by the power companies for various telemetering and power-control functions.

58. Carrier-system Choice. The choice of a carrier system for application between two points depends on a number of factors, the most important of which are naturally the distances concerned, the numbers of channels desired, and the available conductors. The possible optional or joint use of the system for other broad-band needs, such as television, may be involved in the situation.

The distance of transmission between carrier terminals depends on the gain and output capacity provided at the sending terminal, the line-noise conditions, and the receiving gain. Considerable distance can be spanned by the use of amplifiers at the sending end having sufficient gain and load capacity, but there is a practical limit in this connection. Usually it is less expensive to provide moderate power amplifiers at the terminals and employ repeaters at intervals along the route.

Noise into carrier systems, particularly open-wire, may include radio interference, sometimes from stations at some distance and where the frequency allocations of the carrier system and radio stations overlap.

The availability of power to activate the repeaters may, in some instances, be a

practical factor in the choice of a particular carrier system. Some systems, for example, the coaxial system, transmit power to several repeater points over the conductor which carries the speech. Thus they require a basic power supply only at infrequent intervals.

Another factor of considerable importance in laying out a carrier system is the matter of cross talk between systems operating on the same open-wire lead or on different conductors in the same cables. In general, this problem is minimized when systems of the same frequency allocation and power level are operated over the same line structure. Carrier systems tend to be incompatible if the frequency allocations are not the same or closely related. In particular, in the case where the same frequency is employed on nearby conductors but in opposite directions, the cross-talk problem is greatly magnified.

59. Frequency Allocations. In one type of system, often referred to as a *stackable* system, the channels may be added to a pair of wires one at a time. In this case the frequencies are assigned upward from the lowest usable frequency so that a single channel occupies opposite-directional adjacent frequency bands in the available spectrum. A second channel, perhaps installed at a later date, might then be piled on top of the first channel with its opposite-directional frequencies occupying channels 3 and 4. Other added channels would go to higher frequencies, up to a practical limit determined by the line characteristics and the capabilities of the terminal equipment. Such a frequency-allocation arrangement has the advantage of flexibility for growth, but it makes it difficult to apply repeaters at intervals where more than one channel is involved because of the relative complexity of the filter-separation equipment.

The preferred arrangement, where several channels are involved and it is necessary to employ repeaters, uses *grouped frequencies* in which the one-direction bands of several channels are grouped and the other direction of the respective channels occupies a frequency region adjacent to the first group. The two directions of all channels are therefore capable of being separated by a single set of *directional filters* and each passed through a single amplifier. Grouped-frequency systems and stacked-frequency systems are generally incompatible on the same lines because of cross-talk difficulties.

60. Signaling on Carrier Systems. All carrier systems must have provision for signaling and supervision over the channels. Sometimes the signaling arrangements are independent of the carrier system itself and simply consist of an application of signaling systems which are standard for other links in the toll plant. For example, one signaling arrangement applies to the voice band a modulated tone of 2,600 cps. This can be transmitted over a carrier telephone channel which may form part of a long telephone connection. Such a signaling system must, of course, be designed so that the tone is not present when speech is carried over the channel. Correspondingly, the signal-receiving equipment must not be susceptible to being triggered off when voice currents are applied to the sending end.

Another signaling system employs a frequency just outside the active telephone band at 3,700 cps, which is normally not employed for conversation. Such a signaling system is independent of the speech channel and, if necessary, functions with the signaling currents on the line at the same time as speech.

There are other signaling systems which operate by transmitting or interrupting the carrier current itself. It is important in selecting a carrier system to be sure that the signaling system provided is extremely reliable and that it will function under any condition of line attenuation or other influences which will still permit speech to be carried on. At the same time, it must not be triggered off by unwanted noises or transients that may, in some cases, be tolerable in the voice circuit.

61. Some Terminal Details. The channel-separating terminals of carrier systems are frequently also employed to provide multiplexing facilities over broad-band radio circuits. It is therefore convenient to consider in more detail one or two typical carrier-system terminals.

One typical system is the Western Electric *type K 12-channel* carrier system nor-

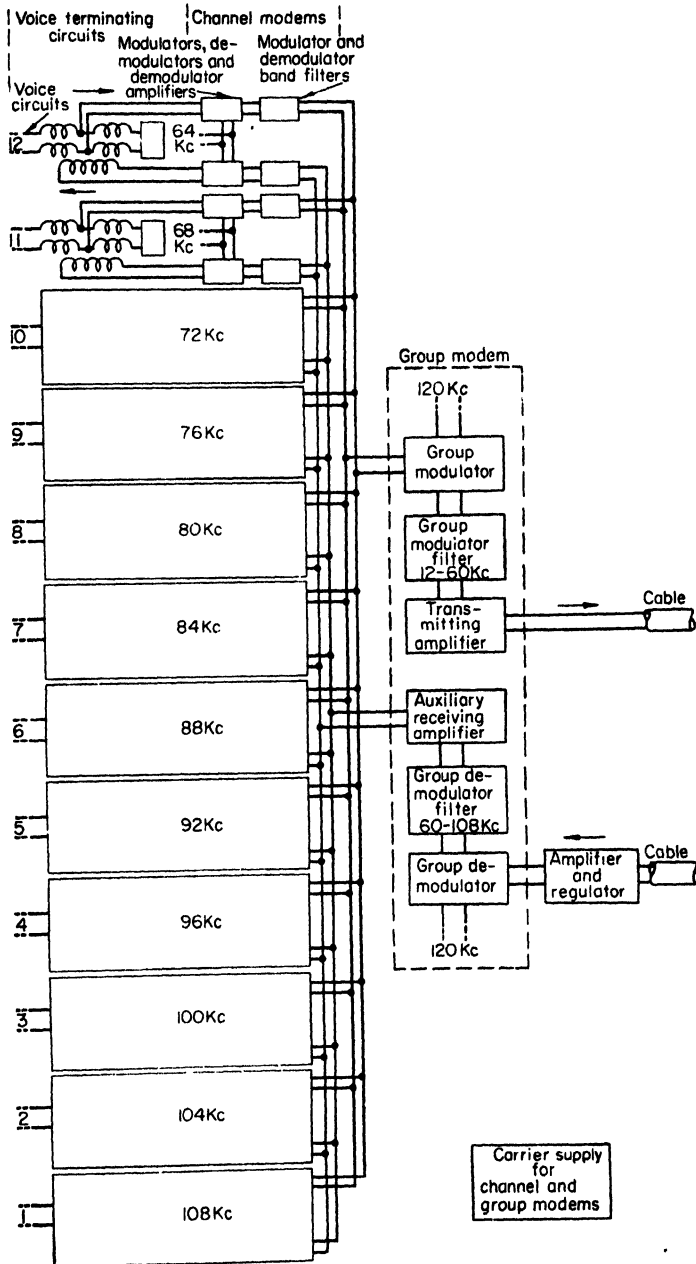



FIG. 34. Type K carrier-system terminal.

mally employed over 19-gage cable conductors, with repeaters spaced at intervals of 17 miles. This system uses vacuum tubes except for the modulators and demodulators, where copper-oxide rectifiers are employed in a balanced arrangement to suppress the carrier. A block diagram of a terminal of this system is shown in Fig. 34.

In this system there is an amplifier common to the 12 channels which is not unlike the amplifier used at repeater points. At the sending terminal of this system arrangements are also made to supply pilot-channel currents at the extremes of the frequency range which serve to regulate the transmission of a long system by changing the gain of the repeaters automatically as the line attenuation varies with temperature.

The channel filters select sharply at intervals of 4,000 cps, and the basic selection is in the frequency range from 60 to 120 kc, at which point very efficient filter operation is achieved using filters which contain quartz crystals as well as the usual inductance-capacitance elements. The frequencies are stepped down to the desired line-frequency range, from 12 to 60 kc, by the use of a second stage of modulation also employing copper-oxide rectifiers.

The K-system terminal is useful as an illustration because the 12-channel so-called *channel bank* is also employed as a building block in the arrangements to provide as many as 1,800 channels in one band for operation over wide-band coaxial systems or microwave radio systems.

Where the 12-channel banks are employed for this purpose, supplementary steps of modulation provide 60-channel *supergroups* which are, in turn, joined to give 600-channel *master groups*. In turn, three master groups can be joined to give 1,800 channels for a modern coaxial system.

This type K channel bank is intended for long-haul operation where many such channels may be operated in tandem for long connections. Thus each channel must have quite precise frequency-transmission characteristics in order that the over-all speech band may not suffer.

62. Frequency Stability Needed. Where large numbers of channels are involved, using very high frequencies, the precision of frequency of the carrier sources of the different channels must be high. This is achieved by deriving all the channel frequencies as well as intermediate modulation steps from a single precise high-stability oscillator.

If carrier frequencies in any system vary, not only will the speech be degraded over the channels because the side bands will not fit into the band filter characteristics, but also in an SSB system if the carrier supply to the demodulator differs from that supplied to the modulator of the corresponding channel, the voice will be displaced upward or downward in the frequency spectrum. This will make it sound unnatural because of the fundamental pitch change and the fact that the harmonics of the voice frequencies are also displaced relative to their natural position. A displacement of 20 cps is quite discernible for ordinary speech. For the transmission of high-grade music, the permissible variation is considerably less.

63. Type P System. Another carrier system of some interest is the Western Electric *type P system,* in which transistors have been exclusively employed for the first time to reduce size and cost and thereby provide a system capable of being justified for comparatively short distances, such as rural subscriber and combination open-wire and cable tributary trunk lines. This is a four-channel system using frequencies up to 100 kc.

A circuit diagram of one type P carrier-system terminal is shown in Fig. 35. It includes, in addition to the regular oscillator-modulator-demodulator functions, a *compandor* for each channel which compresses the volume range of the speech after it passes the modulator. At the corresponding receiving end there is an element using transistors which expands the volume to approximately the same volume range which existed at the sending end.

As a result of this compressor-expandor action, if there is noise on the line circuit, its effect will be reduced since the volumes of the speech-carrier currents on the line are sustained at a higher level above the noise currents than they would otherwise have

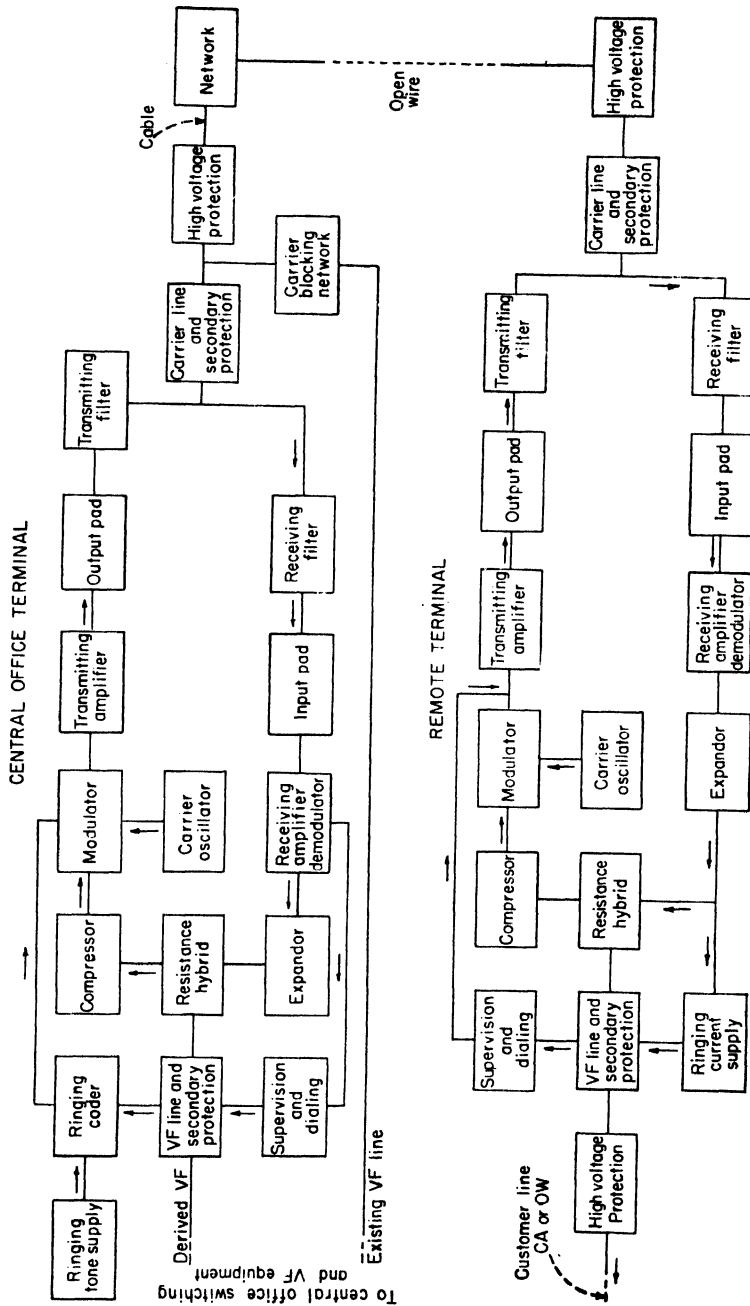



FIG. 35. Over-all block diagram of the P1 carrier system.

been. This advantage approaches 20 db and makes it possible to transmit over a longer distance than would otherwise have been the case for the same system without the compandor. Cross-talk problems are also simplified. The additional complexity of the equipment needed to include the compandor is, in most cases, worth the additional cost which it entails.

Compandors have also been used on other types of carrier systems; in fact, separate compandor units have been designed and standardized for application to other types of telephone circuits where special conditions make it necessary to minimize line noise.

64. Carrier Repeaters. Almost all type K 12-channel cable carrier systems require repeaters in the line circuit. The circuit diagram of a type K repeater is shown in Fig. 36. It is a feedback vacuum-tube device having circuits not unlike the other amplifiers basically, but provided with *pilot-channel* control circuits.

A pilot channel consists of a steady single-frequency current which is transmitted over the line circuit of a carrier system, the frequency being sandwiched between or at the extreme of the speech bands. Since the power of the SSB speech varies continuously, the pilot-channel current being transmitted over the same circuit provides a

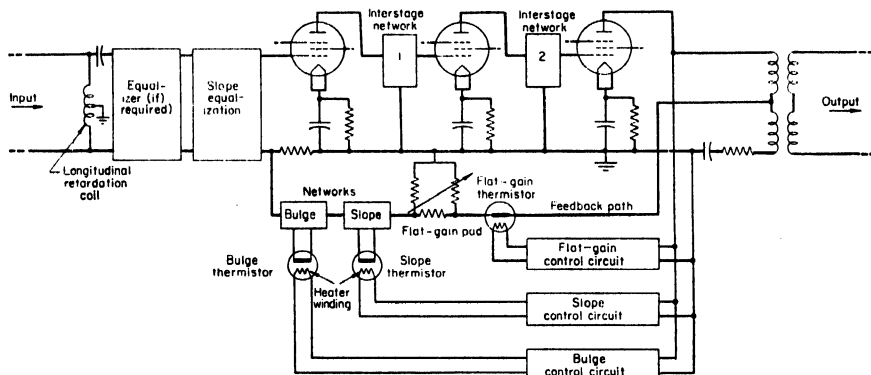



FIG. 36. Type K carrier repeater.

better measure of transmission conditions, and it is affected in transmission like the speech bands. It can be used to adjust automatically the gain of the repeater to compensate for line variations with time, temperature, or other factors. The use of more than one pilot channel makes possible considerable refinement in the smoothness and accuracy of this adjustment over a very wide band.

In the *type K* repeater, the actual control is accomplished by incorporating in the feedback path of the repeater a network which has variable elements, so-called *thermistors*, whose resistances follow the variations since they receive their energy from the pilot currents selectively picked off the output of the repeater.

The design of such automatic control circuits involves some rather nice problems, not only of the static performance of the repeater to match the line variations with frequency and time, but also of the dynamic characteristics of the control circuit. This may be only one element in a complex feedback circuit involving as many as 50 or more other repeaters in tandem in a circuit several hundred miles in length.

As noted previously, the design of the amplifier for a carrier or radio repeater, which has several channels passing through the same set of vacuum tubes or transistors, imposes very severe requirements as to linearity. The intermodulation components must, in general, be kept at least 60 db below the speech channels so that excessive channel cross talk or noise may not result. In general, this calls for the use of a substantial amount of feedback.

The frequency allocation employed can also be a factor in the successful operation of repeaters since it is sometimes possible to arrange that the more serious intermodu-

lation components fall in the intervals between channels. Especially stringent requirements for linearity are imposed when many repeaters are operated in tandem over a long circuit and the resultant intermodulation components add. In designing or choosing a carrier system for long-distance operation, such factors must be carefully considered.

65. L3 Coaxial System. The repeaters used on the Western Electric Company L3 coaxial system deserve a special note. This system provides a total operational band of 8 Mc which can be split by the proper terminal equipment into as many as 2,000 telephone circuits, or used for one 4-Mc television band plus about 800 telephone circuits. It is operated over a standard 0.375-in. coaxial conductor, described previously under Wire Circuits.

Because of the wide frequency band transmitted, repeaters are required every 4 miles. The system is designed for telephone transmission over circuits several thousand miles in length with hundreds of repeaters. This imposes extremely stringent requirements on the transmission design of the individual amplifiers from the standpoint of accuracy of transmission performance, stability with time, and the effectiveness of the multipilot system which is used to control the amplification to make

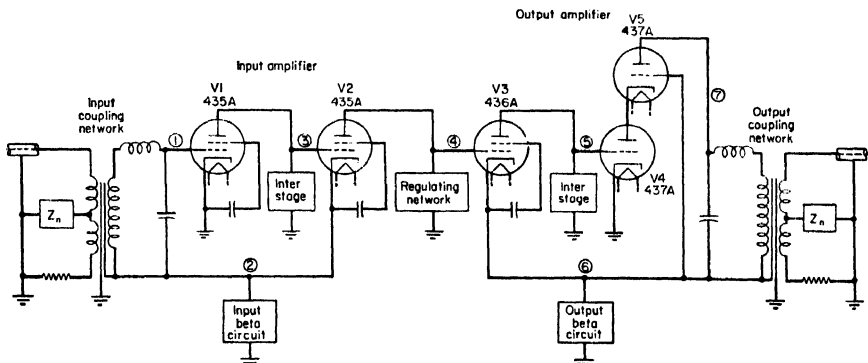



FIG. 37. L3 line-amplifier arrangement.

up for the variation in coaxial conductor attenuation with temperature. This has required, in the design of the system and provisions for its maintenance, much greater attention to detail than perhaps any other type of telephone transmission system ever created.

Because of the large number of repeaters, it has been found necessary to house them in nonattended underground vaults or, preferably, small huts. The repeaters are sealed in airtight containers to minimize the effect of varying moisture conditions. Power to the repeaters is supplied over the coaxial structure itself from a constant-current d-c source with primary power-supply points approximately 75 miles apart. As shown in Fig. 37, the circuit, in effect, consists of two two-stage amplifiers in tandem, each with independent feedback, and the tubes are a special high-transconductance, high-frequency type, designed for long life.

RADIO CIRCUITS

Where radio is used as a link in a telephone system, its transmission requirements should, in general, meet the transmission performance of any other type of circuit employed except where there are extenuating circumstances.

Most radio systems are operated on a multichannel basis, and it is, of course, important that the system be so designed that it will accept speech from the several channels concerned without overloading the radio terminal equipment. It is important, also, that any signaling tones employed do not interfere with other signaling or other speech

Statistics on Some Typical Carrier Systems, United States

Maker	Code	No. of channels	Frequency range, kc	Channel arrangement	Conductors	Repeater spacing, miles	Automatic regulation	Compandor	Signaling	Remarks
Panhandle (Kellogg)	FM	1–10	42.5–342.5	Stackable	Open wire	5–12	Yes	No	Yes	Rural service, no repeaters, f-m transmission
North Electric	RTC-1	1–4	6.7–76.5	Stackable	Open wire	10–22	No	No	Yes	Rural service, no repeaters
Lynch	B-120	1–5	3–62	Stackable	Open wire	11–30	No	No	Yes	Rural service, no repeaters
Lynch	B-109	10	0.2–62	Four-wire	Radio or wire lines	....	Yes	No	No	No repeaters
Lynch	B-37	1–5	3.5–62	Stackable	Open wire	....	Optional	No	Optional	No repeaters
Federal	9H2	1–3	3.6–35	Stackable	Open wire	125–225	No	No	Yes	No repeaters
Lenkurt	32E	3 or 4	3.3–35	Grouped	Open wire	125	Yes	No	Optional	
Lenkurt	33A	1–3	3.5–35	Stackable	Open wire	100–200	Optional	No	Optional	
Lenkurt	33B	2–12	0.2–48	Four-wire	Radio	....	No	No	Optional	11 channels max with signaling
Lenkurt	45A	12	40–150	Grouped	Open wire	30–75	Yes	No	Optional	
Lenkurt	45BN	24	40–264	Grouped four-wire	Multipair cable	7	Yes	Optional	Optional	
Lenkurt	45BX	120	12–528	Four-wire	Radio	....	Yes	Optional	Optional	
Lenkurt	45CB	4	40–76	Grouped	Open wire	70–120	Yes	Yes	Optional	
W. E. Co.	C	3	5–31	Grouped	Open wire	125	Yes	No	No	
W. E. Co.	H	1	4–10	Grouped	Open wire	20–200	No	No	Yes	
W. E. Co.	J	12	36–143	Grouped	Open wire	30–75	Yes	No	No	
W. E. Co.	K	12	12–60	Four-wire	Multipair cable	17	Yes	No	No	
W. E. Co.	L1	600	64–3,100	Four-wire	Coaxial cable	8	Yes	No	No	
W. E. Co.	L3	1,860	308–8,320	Four-wire	Coaxial cable	4	Yes	No	No	
W. E. Co.	M	5	150–425	Grouped	Open wire or power line	50	Yes	No	Yes	Rural service, no repeaters
W. E. Co.	N	12	36–268	Grouped four-wire	Multipair cable	7	Yes	Yes	Yes	
W. E. Co.	O	4–16	2–156	Stackable groups	Open wire	25–100	Yes	Yes	Yes	
W. E. Co.	ON	24	36–268	Grouped four-wire	Multipair cable	7	Yes	Yes	Yes	
W. E. Co.	ON	96	36–548	Four-wire	Radio	....	Yes	Yes	Yes	
W. E. Co.	P	1–4	9–99	Grouped or stackable	Open wire	2–12	Yes	Yes	Yes	Rural service
Stromberg-Carlson	Rural	1–5	24–138	Stackable	Open wire	....	Yes	No	Yes	

channels. If the radio link has repeaters through which several channels pass in common, they must have a sufficiently high degree of linearity.

A substantial number of intercity trunks are provided by microwave-radio systems which are used alternatively or over the same broad-band channel for television. The telephone circuits are operated on a constant-net-loss basis and, in all operational aspects, are indistinguishable from wire circuits which may form part of the same general network.

An exception in operational procedure is the case of the long transoceanic high-frequency radio links. Because of the variable attenuation and the frequent high-noise characteristics of these paths, the circuit is not set up on a constant-net-loss basis. In addition to the automatic volume control, which normally is part of the radio receiver, it has included in it automatic-gain-control devices, so-called VOGADS (voice-operated gain-adjusting device), which control the gain of the transmitting circuit in accordance with the input speech and ensure loading up the transmitting amplifier to essentially full output capacity at all times, thus providing a maximum signal-to-noise ratio. At the same time, this device reduces the gain in the receiving circuit at times of high gain in the transmitter so that the loop circuits from end to end will not sing.

Radio links employed for mobile service to automobiles or ships pose unusually severe transmission problems that call for automatic gain control somewhat similar to the VOGAD at both terminals. In addition, it is often necessary to employ essentially the same frequencies for opposite directions, and this commonly calls for push-button one-way-at-a-time operation. This is difficult, if not impractical, to tie into a large telephone network otherwise operated on a regular two-way basis.

Since most of the essentially radio problems have been considered in some detail elsewhere in this handbook, this radio section is limited in scope to the few comments noted above.

SPECIAL TRANSMISSION CIRCUITS

A telephone system is frequently called on to provide special circuits for different purposes, not necessarily telephony as such. These include simple telegraph transmission, various types of data-transmission systems, program circuits for radiobroadcasting, and broad-band circuits for television-transmission purposes. It seems likely that, in the course of time, more different types of special circuits will be required.

It serves a useful purpose if such special circuit requirements can be satisfied merely by using a normal telephone channel either directly or with minor modification. This is possible if the special uses have requirements, in respect to band width, phase or delay, band-width–energy distribution, and interference-producing or interference-susceptibility characteristics, which do not differ from regular telephone-circuit characteristics. However, this is frequently not the case and the situation requires special treatment, which may call for the design of new equipment or systems to carry out these functions.

For example, for regular radio-program distribution, band widths of at least 5,000, and sometimes up to 8,000 cps or more, are required. When provided over wire circuits, this may call for specially equalized wider-band one-way wire channels with wide-band loading. More frequently, however, it requires the provision for some wider-band channels in the frequency division of the carrier systems which form a large proportion of the longer-haul circuits in the United States. For example, the frequency range otherwise allotted to two or three regular telephone channels may be employed to provide a wider-band channel of 5 to 8 kc or more in width. These arrangements are embodied in special terminal equipment which supplements some of the terminal equipment employed in the carrier systems. It does not, however, in general, call for a major design of the system.

It has long been common practice to use the voice band of an ordinary telephone channel for the transmission of 12 or more carrier telegraph channels. Other forms of data-transmission systems, as well as picture-transmission systems, have been commonly applied to telephone channels. In the latter case particularly, it is necessary

practice to delay-equalize the circuits involved specially because of the sensitivity of the system to this particular kind of distortion.

The transmission of television signals is, of course, a very unique case since it calls for a width of frequency band and precision of amplitude and delay equalization far beyond any ordinary speech channel. Its requirements can be met by the use of special video pairs, as noted under Wire Circuits, for short distances and by coaxial systems, such as the Western Electric Company L3 system and various types of microwave-radio systems.

Selected References

Telephone Sets:

Alexander, T. W., Jr.: Protection of Wire Communication Facilities Serving Power Stations and Substations, *Trans. AIEE*, No. 9, 587; discussion, 591, November, 1953.

Bennett, A. F.: An Improved Circuit for the Telephone Set, *Bell System Tech. J.*, **32**, 611, May, 1953.

Bennett, W.: Telephone-systems Applications of Recorded Machine Announcements, *Trans. AIEE*, No. 8, 478, September, 1953; *Trans. AIEE*, pt. 1, 72, 478, September, 1953.

———: Telephone System Applications of Recorded Machine Announcements, *Elec. Eng.*, **72**, 975, November, 1953.

Bogert, B. P.: On the Band Width of Vowel Formants, *J. Acoust. Soc. Am.*, **25**, 791, July, 1953.

Chase, A. J.: The New Volume Control Telephone, *Bell Lab. Record*, **33**, 361, October, 1955.

Clemency, W. F.: The Speakerphone, *Bell Lab. Record*, 208, June, 1956.

———, F. F. Romanow, and A. F. Rose: The Bell System Speakerphone, *Trans. AIEE*, No. 30, 148, May, 1957.

Cobb, L. J.: Handset for the 500-type Set, *Bell Lab. Record*, **30**, 317, August, 1952.

Emling, J. W.: General Aspects of Hands-free Telephony, *Trans. AIEE*, No. 30, 201, May, 1957.

Fisk, J. B.: Acoustics in Communication, *J. Acoust. Soc. Am.*, **26**, 644, September, 1954.

Gibbon, C. O.: The Common Battery Anti-side Tone Subscriber Set, *Bell System Tech. J.*, **17**, 245, April, 1938.

Hoth, D. F.: Room Noise Spectra at Subscribers' Telephone Locations, *J. Acoust. Soc. Am.*, **12**, 499, April, 1941.

Inglis, A. H.: Transmission Features of the New Telephone Sets, *Bell System Tech. J.*, **17**, 358, July, 1938.

——— and W. L. Tuffnell: An Improved Telephone Set, *Bell System Tech. J.*, **30**, 239, April, 1951.

Martin, W. H.: Seventy-five Years of the Telephone, An Evolution in Technology, *Bell System Tech. J.*, **30**, 214, April, 1951.

Mott, E. E., and R. C. Miner: The Ring Armature Telephone Receiver, *Bell System Tech. J.*, **30**, 110, January, 1951.

New "all-in-one" telephone, that can be used on wall or desk, designed to fit readily any application, has been produced by the Connecticut Telephone & Electric Corporation. The instrument is actually a universal local-battery telephone which can be converted to a manual common-battery telephone without adding either electrical or mechanical parts, *Tel. & Tel. Age*, **70**, 32, December, 1952.

Peterson, G. E.: Applications of Information Theory to Research in Experimental Phonetics; Basic Physical Systems for Communication between Two Individuals, *J. Speech & Hear. Dis.*, **17**, 175, June, 1952; **18**, 116, June, 1953.

Pferd, W.: A Governor for Telephone Dials, *Bell System Tech. J.*, **33**, 2167, November, 1954.

Rickaby, A.: The Loudspeaking Telephone, *A.T.E.J.*, **9**, 28, January, 1953.

Rodek, V.: The Hands-free Telephone, *Trans. AIEE*, No. 30, 210, May, 1957.

Spencer, H. J. C.: Evolution of British Post Office Telephone, *P.O. Telecommun. J.*, **6**, 63, February–April, 1954.

General Switching and Traffic Problems:

Abraham, L. G., A. J. Busch, and F. F. Shipley: Crossbar Toll Switching System, *Trans. AIEE*, **63**, 302, 1944.

Anderson, A. E.: Transistors in Switching Circuits, *Bell System Tech. J.*, **31**, 1207, November, 1952.

Atalla, M. M.: Arcing of Electrical Contacts in Telephone Switching Circuits, Reference, *Bell System Tech. J.*, **33**, 535, May, 1954.

———: Arcing of Electrical Contacts in Telephone Switching Circuits, Part V—Mechanism of the Short Arc and Erosion of Contacts, *Bell System Tech. J.*, **34**, 1081, September, 1955.

Brackett, G. R.: Private Branch Exchange, *Telephony*, **146**, 17, Apr. 10, 1954.

Brown, J. T. L., and C. E. Pollard: Mercury Contact Relays, *Elec. Eng.*, **66**, 1106, November, 1947.

Curtis, A. M.: Contact Phenomena in Telephone Switching Circuits, *Bell System Tech. J.*, **19**, 40, January, 1940.

Ellwood, W. B.: Glass-enclosed Reed Relay, *Elec. Eng.*, **66**, 1104, November, 1947.

Evers, H. G., and F. D. Reese: Line Concentrators for the Telephone Plant, *Telephony*, **151** (9), Mar. 3, 1956.

Fessler, G. E.: A Remote Line Concentrator for No. 5 Crossbar, *Bell Lab. Record*, **35** (5), 168, May, 1957.

Frost, G. R., W. Keister, and A. E. Ritchie: A Throwdown Machine for Telephone Traffic Studies, *Bell System Tech. J.*, **32**, 292, March, 1953.

Joel, A. E., Jr.: Experimental Remote Controlled Line Concentrator, *Bell System Tech. J.*, **35** (2), 249, March, 1956.
Keller, A. C.: New General Purpose Relay for Telephone Switching Systems, *Bell System Tech. J.*, **31**, 1023, November, 1952.
———: New General Purpose Relay, *Bell System Tech. J.*, **31**, 1023, November, 1952; *Trans. AIEE*, **71**, pt. 1, 413, 1952.
———, H. N. Wagar, R. L. Peek, Jr., and M. A. Logan: Design of Relays, *Bell System Tech. J.*, **33** (complete issue), January, 1953.
Morton, K. L.: Time Measurement of Sequential Contacts, *Tele-Tech*, **14** (10), 68, October, 1955.
Rafuse, I. S.: New Multicontact Relay for Telephone Switching Systems, *Bell System Tech. J.*, **33**, 1111, September, 1954.
Reese, F. D.: Use of Line Concentrators, *Telephony*, **151** (17), Oct. 17, 1956.
Riordan, J.: Telephone Traffic Time Averages, *Bell System Tech. J.*, **30**, pt. II, 1129, October, 1951.
Shannon, C. E.: Synthesis of Two-terminal Switching Circuits, *Bell System Tech. J.*, **28**, 59, January, 1949.
Trucksess, D. E.: Regulated Rectifiers in Telephone Offices, *Trans. AIEE*, **61**, 613, 1942.
Truitt, C. J.: Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks, *Bell System Tech. J.*, **33**, 277, March, 1954.
Turner, E. C.: Telephone Growth Forecasting, *Telephony*, **146**, 17, June 12, 1954.
Wilkinson, Roger I.: The Reliability of Holding Measurements, *Bell System Tech. J* **20**, 365, October, 1941.
———: Theories for Toll Traffic Engineering in U.S.A., *Bell System Tech. J.*, **35** (2), 421, 2 charts, March, 1956.
———: Working Curves for Delayed Exponential Calls Served in Random Order, *Bell System Tech. J.*, **32**, 360, March, 1953.

Automatic Switching:

Anderson, A. E.: Transistors in Switching Circuits, *Proc. IRE*, **40**, 1541, November, 1952; Corrections, December, 1952, p. 1732.
Bahler, W. T., J. W. Cohen, and M. M. Jung: Calculation of the Number of First Group Selectors for Telephone Systems Provided with First and Second Concentration Stages, Taking into Account the Internal Blocking, *Commun. News*, **14**, 51, January, 1954.
Barnard, A. J.: Motor Uniselector-type Group Selector for Trunk Switching, *P.O. Elec. Eng. J.*, **46**, pt. 3, 108, October, 1953.
Blashfield, W. H.: Subscriber Toll Dialing Tape Reader, *Trans. AIEE*, No. 5, 17, March, 1953.
Blount, F. E.: Transistor Oscillator for Use in Multi-frequency Pulsing Current Supply, *Bell System Tech. J.*, **32**, 1313, November, 1953.
Brewer, S. T., and G. Hecht: A Telephone Switching Network and Its Electronic Controls, *Bell System Tech. J.*, **34**, 361, March, 1955.
Brown, C. B., and M. E. Terry: The Application of Designed Experiments to the Card Translator, *Bell System Tech. J.*, **33**, 369, March, 1954.
Burkett, A.: General Aspects of Community Dial Office Equipment, *Trans. AIEE*, **69**, pt. 1, 1540, 1950.
Clark, A. B., *et al.*: Automatic Switching for Nationwide Telephone Service, *Trans. AIEE*, **71**, pt. I, 245, September; *Bell System Tech. J.*, **31**, 823, September, 1952.
Clos, C.: Study of Non-blocking Switching Networks, *Bell System Tech. J.*, **32**, 406, March, 1953.
——— and R. I. Wilkinson: Dialing Habits of Telephone Customers, *Bell System Tech. J.*, **31**, 32, January, 1952.
Cohen, J. W.: Some Examples of the Use of Implication in Switching Algebra, *Commun. News*, **16**, 2, October, 1955.
Dahlbom, C. A., A. W. Horton, Jr., and D. L. Moody: Application of Multifrequency Pulsing in Switching, *Trans. AIEE*, **68**, pt. 1, 392, 1949.
Frey, J.: Automatic Telephone System 7AZ (in German), *Tech. Mitt.*, **34**, 92, Feb. 1, 1956.
Gilbert, E. N.: N-terminal Switching Circuits, *Bell System Tech. J.*, **30**, 668, July, 1951.
Hampton, L. N., and J. B. Newsom: The Card Translator for Nationwide Dialing, *Bell System Tech. J.*, **32**, 1037, September, 1953.
Hohn, F. E., and L. R. Schissler: Boolean Matrices and the Design of Combinational Relay Switching Circuits, *Bell System Tech. J.*, **34**, 177, January, 1955.
Joel, A. E.: Electronics in Telephone Switching Systems, *Trans. AIEE*, January, 1957, p. 701.
Johnston, D. F.: Intertoll Trunk Concentrating Equipment, *Bell System Tech. J.*, **33**, 303, March, 1954.
Keister, William: Logic of Relay Circuits, *Trans. AIEE*, **68**, pt. 1, 571, 1949.
King, G. V.: Centralized Automatic Message Accounting System, *Bell System Tech. J.*, **33**, 1331, November, 1954.
Korn, F. A., and J. G. Ferguson: No. 5 Crossbar Dial Telephone Switching System, *Trans. AIEE*, **69**, pt. 1, 244, 1950.
Lee, C. Y.: Analysis of Switching Networks, *Bell System Tech. J.*, **34** (6), 1287, November, 1955.
Letellier, G.: Fully Automatic Toll Telephone Operation (in French), *Echo recherches*, No. 22, 1, January, 1956.
Lewis, W. D.: Electronic Computers and Telephone Switching, *Proc. IRE*, **44**, 1242, October, 1953.
Mallery, P.: Transistors and Their Circuits in the 4A Toll Crossbar Switching System, *Trans. AIEE*, No. 8, 388, September, 1953.
Mealy, G. H.: A Method for Synthesizing Sequential Circuits, *Bell System Tech. J.*, **34**, 1043, September, 1955.

Meszar, John: Fundamentals of the Automatic Telephone Message Accounting System, *Trans. AIEE*, **69**, pt. 1, 255, 1950.
Myers, O.: Common Control Telephone Switching Systems, *Bell System Tech. J.*, **31**, 1086, November, 1952.
Ostline, J. E.: Automatic Call Recording and Accounting in the SATT System, References, *Trans. AIEE*, No. 5, 56, March, 1953.
Ritchie, A. E.: Sequential Aspects of Relay Circuits, *Trans. AIEE*, **68**, pt. 1, 577, 1949.
Rose, W. R.: New Strowger Selector, *G. E. C. Telecomm.*, **6** (2), 51, 1951.
Ruggles, L. L.: Independent Participation in the Plan for Nationwide Customer Toll Dialing, *Tel. Engr.*, **58**, 39, May, 1954.
Schneckloth, H. H.: Some Basic Concepts of Translators and Identifiers Used in Telephone Switching Systems, *Bell System Tech. J.*, **30**, 588, July, 1951.
Shannon, C. E.: Memory Requirements in a Telephone Exchange, *Bell System Tech. J.*, **29**, 343, July, 1950.
Simmons, H. J.: Switching Circuit Techniques, *Tech. Eng. News*, **34**, 20, December, 1952.
Simultaneous Inauguration of Eleven Automatic Telephone Exchanges Employing New Philips System, *Commun. News*, **16**, 37, October, 1955.
Staehler, R. E.: An Application of Boolean Algebra to Switching Circuit Design, *Bell System Tech. J.*, **31**, 280, March, 1952.
Washburn, S. H.: Application of Boolean Algebra to the Design of Electronic Switching Circuits, *Trans. AIEE*, No. 8, 380, September, 1953.

Transmission—General:

Affel, H. A.: Transmitted Frequency Range for Circuits in Broad Band Systems, *Bell System Tech. J.*, **16**, 487, October, 1937.
Bode, H. W.: Feedback Amplifier Design, *Bell System Tech. J.*, **19**, 421, July, 1940.
Carter, C. W., Jr., A. C. Dickieson, and D. Mitchell: Application of Compandors to Telephone Circuits, *Trans. AIEE*, **65**, 1079, 1946.
Caruthers, R. S.: "Levels" and "Powers" in Carrier Systems, *Wire & Radio Commun.*, No. 10, 18, 60, October, 1953.
Chinn, H. A., D. K. Gannett, and R. M. Morris: New Standard Volume Indicator and Reference Level, *Bell System Tech. J.*, **19**, 94, January, 1940.
Clark, A. B.: The Development of Telephony in the United States, *Trans. AIEE*, **71**, pt. 1, 348, 1952.
Coy, J. A.: Heat Dissipation from Toll Transmission Equipment, *Trans. AIEE*, No. 10, 762, January, 1954; *Trans. AIEE*, pt. 1, 72, 762, January, 1954.
Dudley, Homer: Carrier Nature of Speech, *Bell System Tech. J.*, **19**, 495, October, 1940.
Feldman, C. B., and W. R. Bennett: Band Width and Transmission Performance, *Bell System Tech. J.*, **28**, 490, July, 1949.
Fletcher, Harvey: Hearing, the Determining Factor for High-fidelity Transmission, *Proc. IRE*, **30**, 266, June, 1942.
Fowler, N. B.: Measurements in Communications, *Elec. Eng.*, **66**, 135, February, 1947.
Hunt, L. E., and W. J. Albersheim: An Envelope Delay Scanner, *Proc. IRE*, **40**, 454, April, 1952.
Huntley, H. R.: Transmission Design of Intertoll Telephone Trunks, *Bell System Tech. J.*, **32**, 1019, September, 1953.
———: Where We Are and Where We Are Going in Telephone Transmission, *Elec. Eng.*, **75** (8), 686, August, 1956.
Llewellyn, F. B.: Some Fundamental Properties of Transmission Systems, *Proc. IRE*, **40**, 271, March, 1952.
Lovell, C. A., J. H. McGuigan, and O. J. Murphy: An Experimental Polytonic Signaling System, *Bell System Tech. J.*, **34**, 783, July, 1955.
Phillips, G. M.: Echo and Its Effects on the Telephone User, *Bell Lab. Record*, **32** (8), August, 1954.
Riordan, John: A Ladder Network Theorem, *Bell System Tech. J.*, **16**, 303, July, 1937.
Shannon, C. E.: Communication in the Presence of Noise, *Proc. IRE*, **37**, 10, January, 1949.
———: Communication Theory of Secrecy Systems, *Bell System Tech. J.*, **28**, 656, October, 1949.
———: Mathematical Theory of Communication, *Bell System Tech. J.*, **27**, 379, July; 623, October, 1948.

Wire Circuits:

Alexander, T. W., Jr.: Protection of Underground Cable Sheaths, *Iron Steel Engr.*, **30**, 97; discussion, 100, April, 1953.
Babcock, W. C., Miss E. Rentrop, and C. S. Thaeler: "Crosstalk on Open-wire Lines," Bell Telephone System Monograph 2520, 1955.
Blain, R.: New Devices and Techniques, *Telephony*, **146**, 17, Jan. 30, 1954.
———: Pre-spun Cable Installation: Novel Method of Aerial Construction, *Telephony*, **146**, 16, 35, Apr. 3, 1954.
Booth, R. P., and T. M. Odarenko: Crosstalk between Coaxial Conductors in Cable, *Bell System Tech J.*, **19**, 358, July, 1940.
Bryngelson, B. E.: How Case Transposition Brackets Help Expand Toll Circuits, *Telephony*, **146**, 17, June 5, 1954.
Buckley, O. E.: Future of Transoceanic Telephony, *J. IEE*, **89**, pt. 1, 454, 1942.
Chapman, A. G.: Open-wire Crosstalk, Part 1, *Bell System Tech. J.*, **13**, 19, January, 1934.

———: Open-wire Crosstalk, Part 2, *Bell System Tech. J.*, **13**, 163, April, 1934.
Clogston, A. M.: Laminated Conductors for Reducing Skin-effect Losses, *Bell System Tech. J.*, **30**, 491, July, 1951; *Proc. IRE*, **39**, 767, July, 1951.
Coleman, J. O'R., and H. M. Trueblood: Co-ordination of Power and Communication Circuits for Low Frequency Induction, *Trans. AIEE*, **59**, 403, 1940.
Colley, R. H.: The Evaluation of Wood Preservatives, *Bell System Tech. J.*, **32**, 120, January; 425, March, 1953.
———: Poles and Pole Treatment, *Trans. AIEE*, **61**, 685, 1942.
Doba, S., Jr., and A. R. Kolding: A New Local Video Transmission System, *Bell System Tech. J.*, **34** (4), 677, July, 1955.
Drvostep, V. J., and A. W. Lebert: Rigid Coaxial Transmission Lines, *Tele-Tech*, **12**, 78, February, 1953.
Franke, H. C.: Noise Measurements on Telephone Circuits, *Tele-Tech*, **14** (3), 85, March, 1955.
Giese, R. C.: Gas Pressure for Telephone Cables, *Trans. AIEE*, **66**, 471, 1947.
Gilbert, J. J.: A Submarine Telephone Cable with Submerged Repeaters, *Bell System Tech. J.*, **30**, 65, January, 1951.
Gould, K. E.: Crosstalk in Coaxial Cables, *Bell System Tech. J.*, **19**, 341, July, 1940.
Hebbert, C. M.: Toll Cables at Carrier Frequencies, *Bell System Tech. J.*, **20**, 293, July, 1941.
Horn, F. W., and R. B. Ramsey: Bell System Cable Sheath Problems and Designs, *Trans. AIEE*, **70**, pt. 1, 1811, 1951.
Kelly, Mervin J., Sir W. Gordon Radley, G. W. Gilman, and R. J. Halsey: A Transatlantic Telephone Cable, *Elec. Eng.*, **74** (3), 192, March, 1955.
Kine, R. A., and S. P. Morgan: Transmission Formulas and Charts for Laminated Coaxial Cables, *Proc. IRE*, **42**, 1250, August, 1954.
Klein, W.: Historical Development of Cross Talk Theories (in German), Bibliography, *Frequenz*, **8**, 182, June, 1954.
Knox, H. T.: Development of Light Weight Spiral-four Cable, *Wire and Wire Products*, **28**, 467, May, 1953.
Lawson, C. C.: Neoprene Jacketed Telephone Wires, *Wire and Wire Products*, **25**, 829, October, 1950.
Lee, G. W.: Installation of Buried Cables, *Commun. Eng.*, **13**, 33, September–October, 1953.
Leib, F. E.: Copperweld Open Wire for Rural Telephone Construction, *Signal*, **8**, 18, March–April, 1954.
Little, J. S.: Insulation of Telephone Wire with Paper Pulp, *Bell System Tech. J.*, **20**, 82, January, 1941, *Wire and Wire Products*, **14**, 575, October, 1939.
McKay, H.: Radio Interference on Subscriber Lines, *Telephony*, **144**, 46, Apr. 25, 1953.
McRae, J. W., R. F. Mallina, W. P. Mason, T. F. Osmer, and R. H. Van Horn: Solderless Wrapped Connections, *Bell System Tech. J.*, **32**, 523, May, 1953.
Mason, W. P., and O. L. Anderson: Stress Systems in the Solderless Wrapped Connection, *Bell System Tech. J.*, **33**, 1093, September, 1954.
Maupin, J. T.: Measurements in Multipaired Cables, *Bell System Tech. J.*, **30**, 652, July, 1951.
Miller, S. E.: Waveguide as a Communication Medium, *Bell System Tech. J.*, **33**, 1209, November, 1954.
Rounds, P. W., and Miss G. L. Lakin: Equalization of Cables for Local Television Transmission, *Bell System Tech. J.*, **34** (4), 713, July, 1955.
Shaw, T.: Inductive Loading for Bell System Telephone Circuits, *Bell System Tech. J.*, **30**, 149, January; 447, April; 721, July; 1221, October, pt. II, 1951.
Tanner, R. H., and G. B. Thompson: Avoidance of Cross-talk in Large Audio Distribution Systems, *Tele-Tech*, **12**, 80, March, 1953.
Trueblood, H. M., and E. D. Sunde: Lightning Current Observations in Buried Cable, *Bell System Tech. J.*, **28**, 278, April, 1949.
Vaage, E. F.: Transmission-properties of Clogston Type Conductors, *Bell System Tech. J.*, **32**, 695, May, 1953.
Wise, W. H.: Capacity of a Pair of Insulated Wires, *Quart. Applied Math.*, **7**, 432, January, 1950.

Repeaters:

Bardeen, J., and W. H. Brattain: Physical Principles Involved in Transistor Action, *Bell System Tech. J.*, **28**, 239, April, 1949.
Engineers of Automatic Electric Company's Research and Development Laboratories Have Constructed a Series Type Negative-impedance Voice-frequency Repeater for Telephone Communication Believed to Be the First Commercially Available Repeater to Utilize Transistors Instead of Vacuum Tubes, *Fortnightly Tel. Engr.*, **16**, 42, September, 1955.
Merrill, J. L.: The El Telephone Repeater, *Trans. AIEE*, **70**, pt. 1, 49, 1951; *Bell System Tech. J.*, **30**, 88, January, 1951.
———, A. F. Rose, and J. O. Smethurst: Negative Impedance Telephone Repeaters, *Bell System Tech. J.*, **33**, 1055, September, 1954.
Slonczewski, T.: Precise Measurement of Repeater Transmission, *Elec. Eng.*, **73** (4), 346, April, 1954.

Carrier Systems and Related Problems:

Almquist, M. L.: Description of the C-5 Carrier Telephone System, *Proc. Assoc. Am. Railroads*, September, 1940, p. 442.
Appert, K. E., R. S. Caruthers, and W. S. Chaskin: Application and Transmission Features of New 12-channel Open-wire Carrier System, *Trans. AIEE*, **73**, pt. 1 (11), 18, March, 1954.
Barstow, J. M.: Carrier Telephone System for Rural Service, *Trans. AIEE*, **66**, 501, 1947.

Bast, G. H., and D. Goedhart: Carrier System Development Trends, *Commun. News*, **16**, 41, February, 1956.
Bennett, W. R.: Cross-modulation in Multichannel Amplifiers, *Bell System Tech. J.*, **19**, 587, October, 1940.
———: Time Division Multiplex Systems, *Bell System Tech. J.*, **20**, 199, April, 1941.
Black, H. S., F. A. Brooks, A. J. Wier, and I. G. Wilson: Improved Cable Carrier System, *Trans. AIEE*, **66**, 741, 1947.
Boers, Y.: 4-channel Carrier Telephone System for the Argentine VHF Radio Links, *Commun. News.*, **16**, 30, October, 1955.
Boxall, F. S., and R. J. Caruthers: Miniature Compandor for General Use in Wire and Radio Communication Systems, *Trans. AIEE*, **72**, pt. 1 (10), 804, January, 1954.
Boyd, R. C., J. D. Howard, Jr., and L. Pedersen: A New Carrier System for Rural Service, *Bell System Tech. J.*, **36** (2), 349, March, 1957.
Brock, R. L., and R. C. McCarty: On Modulation Levels in Frequency Multiplexed Communication System by Statistical Methods, *Trans. IRE*, II-1 (1), 54, March, 1955.
Caruthers, R. S., W. E. Kahl, and L. Pedersen, The N-1 Carrier Telephone System, *Bell System Tech. J.*, **30**, 1, January; 418, April, 1951.
Chestnut, R. W., L. M. Ilgenfritz, and A. Kenner: Cable Carrier Telephone Terminals, *Bell System Tech. J.*, **17**, 106, January, 1938.
Crane, R. E., J. T. Dixon, and G. H. Huber: Frequency Division Techniques for a Coaxial Cable Network, *Trans. AIEE*, **66**, 1451, 1947.
Dossing, S.: Inter-channel Interference in Multichannel Carrier Telephone Systems, References, *Telecommun. J.*, *Australia*, **9**, 101, October, 1952.
Eberhart, E. K., F. J. Hallenbeck, and E. H. Perkins: Circuit and Equipment Descriptions of Type-Pl Carrier System, *Trans. AIEE*, No. 24, 195, May, 1956.
Edwards, P. G., and L. R. Montfort: The Type-O Carrier System, *Bell System Tech. J.*, **31**, 688, July, 1952.
Elmendorf, C. H., and others: L_3 Coaxial System, References, *Trans. AIEE*, No. 8, 395, September, 1953.
Ford, G. T., and E. J. Walsh: The Development of Electron Tubes for a New Coaxial System, *Bell System Tech. J.*, **30**, pt. II, 1103, October, 1951.
——— and ———: High-transconductance Tubes for Broad-band Telephone System Uses, *Proc. Natl. Elec. Conf.*, **7**, 304, 1951.
Fracassi, R. D., and H. Kahl: Type-ON Carrier Telephone, *Trans. AIEE*, **72**, pt. 1 (10), 713, January, 1954.
Goodall, W. M.: Telephony by Pulse Code Modulation, *Bell System Tech. J.*, **26**, 395, July, 1947.
Gould, K. E.: Equalization of Coaxial Lines, *Trans. AIEE*, **68**, pt. 2, 1187, 1949.
Hearn, R. B., and L. A. Weber: Bl Alarm and Control System for Remote Control of Coaxial Cable Stations, *Trans. AIEE*, pt. 1, 622, 1949.
Holloway, D. G.: New Carrier Equipment for Short Lines. The "Q" System, *Strowger J.*, **8**, 163, November, 1952.
Huntley, H. R.: Factors in the Selection of Telephone Carrier Systems, *Telephony*, **147**, 40, Sept. 25, 1954.
———: Factors in the Selection of Telephone Carrier Systems, II, *Telephony*, **147**, 26, Oct. 2, 1954.
Ingram, S. B.: Decade of Progress in Use of Electronic Tubes. Part I—Communication, *Trans. AIEE*, **59**, 643, 1940.
Ketchledge, R. W., and T. R. Finch: L_3 Coaxial System—Equalization and Regulation, References, *Trans. AIEE*, No. 8, 414, September, 1953.
Kinzer, J. P.: Stability of Tandem Regulators in the L-1 Carrier System, *Trans. AIEE*, **68**, pt. 2, 1179, 1949.
The L_3 Coaxial System, *Bell System Tech. J.*, **32**, 779, July, 1953, complete issue—six papers.
Lawton, J.: Compandors for Reducing Noise and Crosstalk, *A.T.E. J.*, **9**, 109, April, 1953.
Lozier, J. C.: Instantaneous Compandors on Speech Channels, *Bell System Tech. J.*, **30**, pt. II, 1214, October, 1951.
Meacham, L. A., and E. Peterson: Experimental Multichannel Pulse Code Modulation System of Toll Quality, *Bell System Tech. J.*, **27**, 1, January, 1948.
Morrison, L. W., Jr.: Television Terminals for Coaxial Systems, *Trans. AIEE*, **68**, pt. 1, 1193, 1949.
Moynihan, J. D.: Noise Limits Power-line Carrier, *Elec. World*, **141** (6), 85, Feb. 8, 1954.
Newell, N. A., and A. Weaver: Single-frequency Signaling System for Long Telephone Trunks, *Trans. AIEE*, **70**, pt. 1, 487, 1951.
Nyquist, H., and K. W. Pfleger: Effect of Quadrature Component in Single Sideband Transmission, *Bell System Tech. J.*, **19**, 63, January, 1940.
Oliver, B. M.: Automatic Volume Control as a Feedback Problem, *Proc. IRE*, **36**, 466, April, 1948.
Perkins, E. H., and J. J. Mahoney: Type-N Carrier Telephone Deviation Regulator, *Trans. AIEE*, **73**, pt. 1, 757, January, 1954.
Smith, W. T., and J. M. Flanigan: Use of Subscriber Line Carrier in Design of Rural Telephone Systems, *Telephony*, **149**, 19, Nov. 12, 1955.
Snook, J. M.: Expanding with Radio and Carrier, *Tel. Engr.*, **57**, 40, May, 1953.
Spencer, H. H.: Power Supplies for Coaxial Systems, *Trans. AIEE*, **68**, pt. 1, 627, 1949.
Tidd, W. H., S. Rosen, and H. A. Wenk: Test Sets and Methods for Cable Carrier Systems, *Trans. AIEE*, **66**, 726, 1947.
Wallenstein, G. D.: Power and Level in Carrier Systems, *Tel. Engr.*, February, 1954, p. 58.

Weaver, A., and N. A. Newell: In-band Single-frequency Signaling, *Bell System Tech. J.*, **33**, 1309, November, 1954.

Radiotelephone Systems:

Black, H. S., J. W. Beyer, T. J. Grieser, and F. A. Polkinghorn: Multi-channel Microwave Relay System, *Trans. AIEE*, **65**, 798, 1946.
Dorff, L. A.: Operational Study of a Highway Mobile Telephone System, *Trans. AIEE*, **70**, pt. 1, 31, 1951.
Grieser, T. J., and A. C. Peterson: A Broad-band Transcontinental Radio Relay System, *Elec. Eng.*, **70**, 810, September, 1951.
Heising, R. A.: Radio Extension Links to the Telephone System, *Bell System Tech. J.*, **19**, 611, October, 1940.
Klenk, L. M., *et al.*: Multichannel Single-sideband Radio Transmitter Design, *Proc. IRE*, **40**, 783, July, 1952.
Meyers, S. T.: Nonlinearity in Frequency-modulation Radio Systems Due to Multipath Propagation, *Proc. IRE*, **34**, 256, May, 1946.
Monk, N.: Experimental Radio-telephone Service for Train Passengers, *Proc. IRE*, **39**, 873, August, 1951.
Polkinghorn, F. A.: Commercial Single Sideband Radio-telephone Systems, *Communications*, **28**, 24, December, 1948.
Roetken, A. A., K. D. Smith, and R. W. Friis: The TD-2 Microwave Radio Relay System, *Bell System Tech. J.*, **30**, pt. II, 1041, October, 1951.

Special Transmission Circuits:

Cowan, F. A., R. G. McCurdy, and I. E. Lattimer: Engineering Requirements for Program Transmission Circuits, *Bell System Tech. J.*, **20**, 235, April, 1941; *Trans. AIEE*, **60**, 142, 1941.
Dagnall, C. H., and P. W. Rounds: Delay Equalization of Eight-kilocycle Carrier Program Circuits, *Bell System Tech. J.*, **28**, 181, April, 1949.
Doba, S., Jr., and J. W. Rieke: Clampers in Video Transmission, *Trans. AIEE*, **69**, pt. 1, 477, 1950.
Farkas, F. S., F. J. Hallenbeck, and F. E. Stehlik: Band Pass Filter, Band Elimination Filter and Phase Simulating Network for Carrier Program Systems, *Bell System Tech. J.*, **28**, 196, April, 1949.
Leconte, R. A., D. B. Penick, C. W. Schramm, and A. J. Wier: Carrier System for 8000-cycle Program Transmission, *Bell System Tech. J.*, **28**, 165, April, 1949.
Strieby, M. E., and J. F. Wentz: Television Transmission over Wire Lines, *Bell System Tech. J.*, **20**, 62, January, 1941.

INDEX

A

INDEX

A

B

M

N

U

V

W

Y

Z